

Thông tin - Thư viện



1162837

ĐỒ TRUNG TUẦN

PHÂN TÍCH THỐNG KÊ VÀ KHAI PHÁ DỮ LIỆU



TỦ SÁCH KHOA HỌC
MS: 254-KHTN-2018



NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA HÀ NỘI

**PHÂN TÍCH THỐNG KÊ
VÀ KHAI PHÁ DỮ LIỆU**

Lương Khắc Định

Đỗ Trung Tuấn

PHÂN TÍCH THỐNG KÊ VÀ KHAI PHÁ DỮ LIỆU

NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA HÀ NỘI

Lời nói đầu.....	15
------------------	----

Chương 1

Phát triển về phân tích thống kê và khai phá dữ liệu

1.1.	Nền tảng của thực hành khai phá dữ liệu	20
1.1.1.	Lịch sử của khai phá dữ liệu và thống kê.....	21
1.1.2.	Thống kê hiện đại.....	22
1.1.3.	Hai góc nhìn về thực tại	26
1.1.4.	Phân tích thống kê hiện đại: thể hệ hai	28
1.1.5.	Các phương pháp học máy: thể hệ ba	30
1.1.6.	Lí thuyết học thống kê: thể hệ thứ tư	31
1.2.	Khía cạnh lí thuyết đối với khai phá dữ liệu.....	33
1.2.1.	Phương pháp khoa học	34
1.2.2.	Một số khái niệm liên quan đến khai phá dữ liệu.....	35
1.2.3.	Khung lí thuyết cho quá trình khai phá dữ liệu	37
1.2.4.	Tính chắc chắn của quá trình khai phá dữ liệu.....	39
1.2.5.	Dữ liệu hướng người dùng so với dữ liệu hướng tính toán.....	39
1.2.6.	Thay đổi khung dữ liệu.....	42
1.2.7.	Các hoạt động chính của khai phá dữ liệu	43
1.2.8.	Thách thức chính của khai phá dữ liệu.....	46
1.2.9.	Các khía cạnh chính trong khai phá dữ liệu	47

1.2.10.	Yêu cầu chung để thành công trong dự án khai phá dữ liệu.....	49
1.3.	Một số chủ đề liên quan đến khai phá dữ liệu.....	51
1.3.1.	Mẫu hành động	53
1.3.2.	Trực giác của người	53
1.3.3.	Bài toán khai phá dữ liệu	56
1.4.	Yêu cầu đối với dữ liệu trong bài toán thí dụ	60
1.4.1.	Các nhiệm vụ cần thực hiện	60
1.4.2.	Các loại tri thức được khám phá.....	61
1.4.3.	Dữ liệu liên quan tới nhiệm vụ	64
1.4.4.	Phân cấp khái niệm.....	68
1.4.5.	Độ đo về tiêu chí quan tâm.....	73
1.4.6.	Thể hiện và trực quan hóa các mẫu được phát hiện	75
1.5.	Ngôn ngữ hỏi dữ liệu	77
1.5.1.	Vai trò của ngôn ngữ hỏi trong khai phá dữ liệu	77
1.5.2.	Về ngôn ngữ DML	80
1.6.	Kết luận.....	82

Chương 2

Dữ liệu để phân tích thống kê và khai phá

2.1.	Hiểu dữ liệu	83
2.1.1.	Một số định nghĩa	83
2.1.2.	Khía cạnh cần quan tâm về hiểu dữ liệu	85
2.1.3.	Khía cạnh cần quan tâm về chuẩn bị dữ liệu	85
2.2.	Một số nhiệm vụ đơn lẻ về chuẩn bị dữ liệu	85
2.2.1.	Thu thập dữ liệu.....	85
2.2.2.	Trích dữ liệu.....	86
2.2.4.	Mô tả dữ liệu.....	88
2.2.5.	Đánh giá dữ liệu.....	89

2.2.6.	Trắc diện dữ liệu	89
2.2.7.	Làm sạch dữ liệu	89
2.2.8.	Chuyển đổi dữ liệu	90
2.2.9.	Thay thế dữ liệu	92
2.2.10.	Trọng số dữ liệu và cân đối	98
2.2.11.	Lọc dữ liệu và làm tròn	99
2.2.12.	Tóm lược dữ liệu	101
2.2.13.	Rút gọn dữ liệu	104
2.2.14.	Lấy mẫu dữ liệu	104
2.2.15.	Rời rạc hóa các giá trị	109
2.2.16.	Suy diễn dữ liệu	109
2.3.	Độ đo dữ liệu	110
2.3.1.	Giới thiệu về độ đo	110
2.3.2.	Các kiểu đo	111
2.3.3.	Đo khoảng cách	113
2.3.4.	Chuyển dữ liệu	117
2.3.5.	Dạng dữ liệu	117
2.3.6.	Chất lượng dữ liệu đối với phép đo riêng	119
2.3.7.	Chất lượng dữ liệu đối với việc thu thập dữ liệu	121
2.4.	Hiển thị và thăm dò dữ liệu	123
2.4.1.	Giới thiệu về hiển thị	123
2.4.2.	Thí dụ về tóm tắt dữ liệu	123
2.4.3.	Các công cụ hiển thị các biến đơn	125
2.4.4.	Các công cụ hiển thị mối quan hệ giữa hai biến	126
2.4.5.	Các công cụ hiển thị trên quá hai biến	126
2.4.6.	Phân tích thành phần chính	127
2.4.7.	Thang đo nhiều chiều	130
2.4.8.	Đánh giá hiệu năng	134

2.5.	Phân phối dữ liệu.....	146
2.5.1.	Giới thiệu về phân phối dữ liệu.....	147
2.5.2.	Một số thuật ngữ.....	149
2.5.3.	Hàm phân phối tích lũy	151
2.5.4.	Phân phối xác suất rời rạc.....	152
2.5.5.	Phân phối xác suất liên tục	154
2.5.6.	Định nghĩa Kolmogorov	155
2.5.7.	Tạo số ngẫu nhiên	156
2.5.8.	Các ứng dụng.....	156
2.5.9.	Phân phối xác suất thông dụng.....	156
2.5.10.	Công dụng khi kết hợp phân phối tiên nghiệm trong suy luận Bayes.....	159
2.5.11.	Một số phân phối một chiều quen thuộc	160
2.5.12.	Một số đặc trưng của đại lượng ngẫu nhiên	166
2.6.	Góc nhìn về phân tích dữ liệu	169
2.6.1.	Ma trận dữ liệu	169
2.6.2.	Thuộc tính	170
2.6.3.	Khía cạnh đại số và góc nhìn hình học về dữ liệu.....	172
2.6.4.	Góc nhìn xác suất đôi với dữ liệu	177
2.7.	Quá trình chuẩn bị dữ liệu.....	187
2.7.1.	Quá trình khám phá dữ liệu	187
2.7.2.	Khai phá dữ liệu, mô hình hóa và các công cụ mô hình hóa.....	195
2.8.	Bản chất của thế giới và tác động của nó với việc chuẩn bị dữ liệu.....	199
2.8.1.	Đo về thế giới.....	199
2.8.2.	Các loại số đo	201
2.8.3.	Mở rộng đối với thuộc tính của biến.....	203
2.8.4.	Xây dựng các dạng biểu diễn của dữ liệu, tiện cho khám phá.....	205

2.9.	Quá trình chuẩn bị dữ liệu.....	208
2.9.1.	Bước 1 về chuẩn bị dữ liệu.....	209
2.9.2.	Bước 2 về khảo sát dữ liệu	210
2.9.3.	Bước 3 về mô hình hóa dữ liệu	211
2.9.4.	Sử dụng mô hình.....	211
2.9.5.	Công cụ mô hình hóa và chuẩn bị dữ liệu.....	212
2.9.6.	Các giai đoạn chuẩn bị dữ liệu.....	215
2.10.	Kết luận	220

Chương 3

Phương pháp, công cụ phân tích thống kê và khai phá dữ liệu

3.1.	Dữ liệu đặc trưng	222
3.1.1.	Đặc trưng.....	222
3.1.2.	Biến đặc trưng.....	222
3.1.3.	Phương pháp xếp hạng đặc trưng	223
3.1.4.	Phương pháp lựa chọn tập con	226
3.2.	Phương pháp chọn mẫu	229
3.2.1.	Phương pháp chọn mẫu trong nghiên cứu khoa học	229
3.2.2.	Giới thiệu phương pháp khởi động.....	235
3.2.3.	Giới thiệu các phương pháp nghiên cứu định tính.....	237
3.3.	Xử lý phân tích trực tuyến.....	247
3.3.1.	Việc sử dụng OLAP	247
3.3.2.	Vai trò của cơ sở dữ liệu đối với OLAP	247
3.3.3.	Những hoạt động trực tuyến.....	249
3.3.4.	Hỏi bằng ngôn ngữ SQL	250
3.3.5.	Các công cụ OLAP và tính chất.....	250
3.3.6.	Các loại xử lý của người phân tích	251

3.4.	Một số lỗi khi khai phá dữ liệu	255
3.4.1.	Giới thiệu về sai sót.....	256
3.4.2.	Thiếu dữ liệu.....	256
3.4.3.	Chú trọng vào huấn luyện	257
3.4.4.	Chỉ tin cậy một kĩ thuật.....	259
3.4.5.	Đặt câu hỏi sai.....	260
3.4.6.	Chỉ tin vào dữ liệu	261
3.4.7.	Chấp nhận dò rỉ trong tương lai	265
3.4.8.	Giảm trường hợp phiền phức	265
3.4.9.	Ngoại suy	266
3.4.10.	Đáp ứng mọi yêu cầu	267
3.4.11.	Mẫu ngẫu nhiên	268
3.4.12.	Tin mô hình tốt nhất	269
3.4.13.	Định hướng	269
3.5.	Kết luận	270

Chương 4

Một số kĩ thuật phân tích dữ liệu

4.1.	Các thuộc tính số	273
4.1.1.	Phân tích một biến	273
4.1.2.	Phân tích hai biến.....	280
4.1.3.	Phân tích đa biến.....	286
4.1.4.	Chuẩn hóa dữ liệu.....	290
4.1.5.	Phân phối chuẩn.....	292
4.2.	Các thuộc tính phạm trù.....	297
4.2.1.	Phân tích một biến	298
4.2.2.	Phân tích hai biến.....	306
4.2.3.	Phân tích đa biến.....	315
4.2.4.	Khoảng cách và góc	320
4.2.5.	Phân biệt.....	322

4.3.	Dữ liệu đồ thị.....	324
4.3.1.	Khái niệm về đồ thị.....	324
4.3.2.	Các thuộc tính hình dạng.....	329
4.3.3.	Phân tích hướng tâm	334
4.3.4.	Các mô hình đồ thị.....	344
4.4.	Dữ liệu chuỗi thời gian	362
4.4.1.	Giới thiệu về chuỗi thời gian	362
4.4.2.	Trung bình động	366
4.4.3.	Phân tích xu thế.....	370
4.4.4.	Các mô hình ARMA và ARIMA (Box - Jenkins).....	374
4.4.5.	Phân tích phổ	385
4.5.	Kết luận	387

Chương 5

Mô hình phân tích và khai phá dữ liệu

5.1.	Các mô hình và khuôn mẫu	389
5.1.1.	Giới thiệu về mô hình và khuôn mẫu	389
5.1.2.	Cơ bản của mô hình hóa	390
5.1.3.	Các cấu trúc mô hình dự đoán	390
5.1.4.	Các mô hình về phân phối xác suất và các hàm mật độ.....	394
5.1.5.	Khắc phục số chiều	398
5.1.6.	Các mô hình về cấu trúc dữ liệu	399
5.1.7.	Các cấu trúc khuôn mẫu	400
5.2.	Độ phức tạp của mô hình.....	403
5.2.1.	Giới thiệu về độ phức tạp của mô hình	403
5.2.2.	Tập mô hình.....	404
5.2.3.	Độ phức tạp.....	407
5.2.4.	Độ tự do tổng quát.....	409
5.2.5.	Thí dụ về bề mặt có ổn	410
5.2.6.	Nhận xét về mô hình	410

5.3.	Mô hình phù hợp với mục đích.....	415
5.3.1.	Giới thiệu về tính phù hợp	415
5.3.2.	Bài học từ tự nhiên và kĩ nghệ.....	415
5.3.3.	Ra quyết định trong tổ chức kinh doanh.....	417
5.3.4.	Luật hành động	421
5.3.5.	Thí dụ về mô hình hóa nhanh	421
5.4.	Phân tích thống kê của dữ liệu bình thường và bất thường	424
5.4.1.	Giới thiệu về phân tích thống kê.....	424
5.4.2.	Các biểu đồ điều khiển một biến	425
5.4.3.	Các biểu đồ điều khiển nhiều biến	454
5.4.4.	Nhận xét về xử lí dữ liệu bất thường	458
5.5.	Phân tích dữ liệu và không chắc chắn	459
5.5.1.	Giới thiệu dữ liệu không chắc chắn.....	459
5.5.2.	Xử lí không chắc chắn.....	459
5.5.3.	Các biến ngẫu nhiên và các mối quan hệ.....	462
5.5.4.	Các mẫu và suy luận thống kê	465
5.5.5.	Ước lượng.....	467
5.5.6.	Kiểm định giả thuyết thống kê.....	475
5.5.7.	Các phương pháp lấy mẫu.....	481
5.5.8.	Nhận xét về độ chắc chắn	485
5.6.	Kết luận	485

Chương 6

Khai phá luật kết hợp

6.1.	Tổng quan về khai phá dữ liệu	488
6.1.1.	Giới thiệu về khai phá dữ liệu.....	488
6.1.2.	Bản chất của bộ dữ liệu	489
6.1.3.	Các loại cấu trúc theo mô hình và khuôn mẫu	490

6.1.4.	Các nhiệm vụ khai phá dữ liệu	491
6.1.5.	Các thành phần của thuật toán khai phá dữ liệu	494
6.1.6.	Vai trò tương tác trong phân tích thống kê và khai phá dữ liệu	496
6.1.7.	Thuật ngữ trong quá trình khai phá dữ liệu	497
6.2.	Về các thuật toán khai phá dữ liệu theo quan điểm hệ thống	498
6.2.1.	Giới thiệu về thuật toán.....	498
6.2.2.	Thí dụ về thuật toán CART xây dựng bộ phân lớp theo cây.....	501
6.2.3.	Nhìn nhận sơ bộ về các thuật toán khai phá dữ liệu	505
6.2.4.	Nhận xét chung về thuật toán khai phá dữ liệu	514
6.3.	Khai phá luật kết hợp	515
6.3.1.	Nhu cầu về luật kết hợp	515
6.3.2.	Các phương pháp khai phá tập mục phổ biến.....	521
6.3.3.	Các phương pháp đánh giá mẫu	540
6.3.4.	Kết luận về khai phá luật kết hợp.....	548
6.4.	Kết luận	550

Chương 7

Phân lớp và phân cụm

7.1.	Về phân lớp và phân cụm	551
7.2.	Phân lớp	551
7.2.1.	Khái niệm cơ bản.....	552
7.2.2.	Cây quyết định	555
7.2.3.	Các phương pháp phân lớp Bayes.....	573
7.2.4.	Bộ phân lớp dựa trên luật	578
7.2.5.	Đánh giá và lựa chọn mô hình	585
7.2.6.	Các kĩ thuật cải thiện độ chuẩn xác phân lớp	599
7.2.7.	Kết luận về phân lớp.....	607

7.3.	Phân tích cụm.....	609
7.3.1.	Bài toán phân cụm.....	609
7.3.2.	Các phương pháp phân hoạch	617
7.3.3.	Các phương pháp phân cấp.....	624
7.3.4.	Các phương pháp dựa trên mật độ	637
7.3.5.	Các phương pháp dựa vào lưới	646
7.3.6.	Đánh giá việc phân cụm.....	652
7.3.7.	Kết luận về phân cụm.....	660
7.4.	Kết luận	661
	Tài liệu tham khảo.....	663
	Index.....	665

Giáo trình này được sử dụng như tài liệu về chuyên đề *Phân tích thống kê và khai phá dữ liệu*, hay *Khai phá dữ liệu* cho lớp sau đại học về công nghệ thông tin. Trước hết, do nhu cầu của chuyên ngành; sau là cung cấp tài liệu cô đọng, trong bối cảnh có rất nhiều tài liệu tiếng nước ngoài về khai phá dữ liệu; và cuối cùng là do việc cập nhật chương trình đào tạo để phù hợp với trình độ của các nước tiên tiến về công nghệ thông tin và truyền thông.

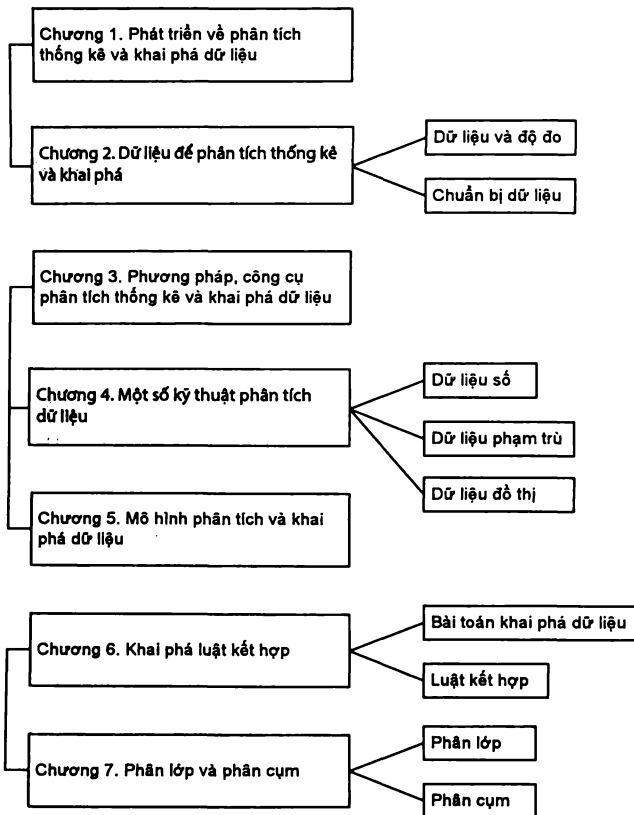
Để có tài liệu cho quá trình đào tạo, ngoài tài liệu bằng tiếng nước ngoài, học viên vẫn cần tài liệu tham khảo bằng tiếng Việt. Các chương trong cuốn sách này nhằm mục đích đó. Tuy nội dung các chương đều tham khảo từ nhiều tài liệu của các cơ sở đào tạo tiên tiến nước ngoài, bản giới thiệu lần đầu này không thể tránh khỏi một số thiếu sót, vẫn cần được chỉnh lí tiếp.

Liên quan đến thuật ngữ sử dụng trong cơ sở dữ liệu, do tiến hóa nhanh của công nghệ cơ sở dữ liệu, một số khái niệm mới và chưa thật thông dụng sẽ được Việt hóa, với chú thích tiếng Anh ở cuối trang. Một số khái niệm có trong định nghĩa thường được ghi chú bên cạnh bằng tiếng Anh để tiện đối chiếu.

Các chương được đánh số theo quy định tiện cho xuất bản, theo chữ số Ả Rập. Một số chú thích ngay trong trang là cần thiết, thay vì trích dẫn trong danh sách các tài liệu tham khảo.

Về nội dung, giáo trình gồm có bảy chương, về tổng thể có thể xem mục đích của các chương đầu nhằm giới thiệu về dữ liệu và các phương pháp thống kê xử lí dữ liệu; các chương sau nhằm trình bày các phương pháp khai phá dữ liệu. Các chương đầu liên quan đến (i) dữ liệu; (ii) phương pháp xử lí dữ liệu theo thống kê; và (iii) mô hình phân tích và khai phá dữ liệu. Các chương sau liên

quan đến (i) luật kết hợp; (ii) phân lớp; và (iii) phân cụm. Chương đầu đã được đóng góp để chỉnh lí, không lệ thuộc vào tài liệu tham khảo về triết học của thống kê và khai phá dữ liệu.



Các chương của giáo trình và nội dung chính

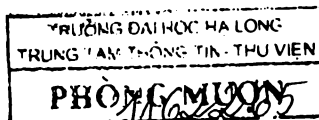
Do giáo trình đề cập nhiều chủ đề, nên đối với mỗi chủ đề, kiến thức chưa được sâu. Học viên có thể sử dụng tài liệu trích dẫn để tìm hiểu thêm về thuật toán hay các nghiên cứu sâu hơn. Cũng có thể sau mỗi chương có tài liệu tham chiếu riêng, nhưng số lượng không nhiều, nên tạm thời các tài liệu tham chiếu được xếp ở cuối sách. Một số khái niệm về thống kê, chọn mẫu dữ

liệu... là khái niệm trong các tài liệu chuyên đề khác, nhưng cần thiết đối với sinh viên và học viên hiểu rõ khái niệm về phân tích thống kê và khai phá dữ liệu, nên được giới thiệu lại trong nội dung giáo trình.

Giáo trình dùng cho sinh viên và học viên tham khảo, không có các bài tập hay hướng dẫn thực hành cuối các chương. Các chương được trình bày không hoàn toàn theo thứ tự phát triển của tri thức, mà chúng đề cập các khía cạnh khác nhau trong phân tích thống kê và khai phá dữ liệu. Do vậy, bạn đọc không nhất thiết truy cập lần lượt đến các chương, mà có thể dùng nội dung các chương với mục đích riêng lẻ.

Mặc dù tác giả đã có nhiều cố gắng trong quá trình biên soạn nhưng chắc rằng cuốn sách vẫn còn những thiếu sót. Tác giả xin trân trọng cảm ơn mọi ý kiến đóng góp, xây dựng từ phía bạn đọc.

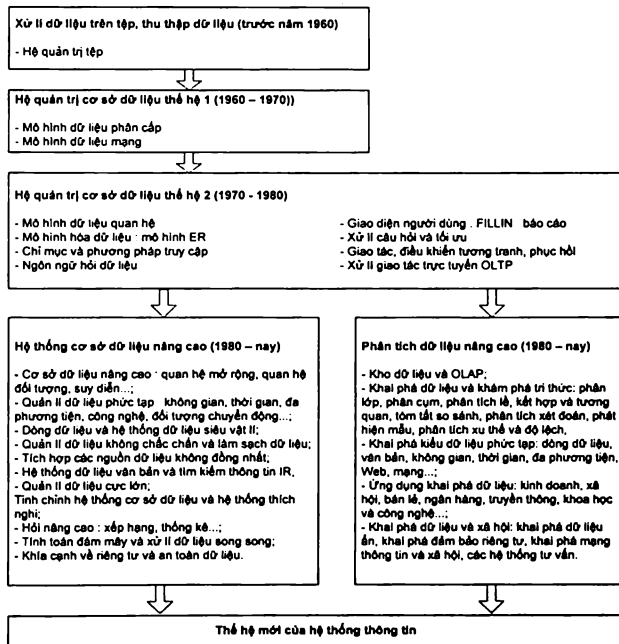
Tác giả



Chương 1

Phát triển về phân tích thống kê và khai phá dữ liệu

Khoa học về dữ liệu bắt đầu từ việc đếm, xử lý thủ công. Trên máy tính, dữ liệu được tổ chức trên tệp, rồi được xử lý tự động. Phát triển của hệ thống cơ sở dữ liệu ảnh hưởng lớn đến phân tích thống kê và khai phá dữ liệu. Hình sau cho thấy một số mốc tiến hóa của các công nghệ sử dụng trong các hệ thống cơ sở dữ liệu.



Hình 1.1. Tiến hóa của công nghệ hệ thống cơ sở dữ liệu¹

¹ FILLIN: fill in form, giao diện người dùng kiểu khuôn dạng.
OLTP: Online Transaction Processing: xử lý giao tác trực tuyến.

Hình 1.1 cho thấy những mốc chính trong quá trình phát triển của các hệ thống cơ sở dữ liệu; trong đó có những mốc liên quan đến phân tích thống kê và khai phá dữ liệu. Chương này trình bày các mốc lịch sử và nền tảng lí thuyết đối với phân tích thống kê và khai phá dữ liệu, và tích hợp chúng với việc phát hiện dữ liệu và chuẩn bị dữ liệu. Những khái niệm này sẽ cần thiết cho việc mô hình hóa dữ liệu.

1.1. Nền tảng của thực hành khai phá dữ liệu

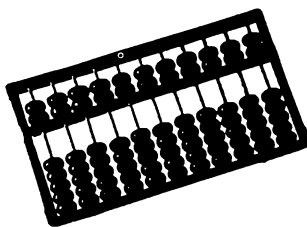
Người ta thấy có nhiều tài liệu về lĩnh vực khai phá dữ liệu; kiến thức về khai phá dữ liệu là rộng, do (i) nó có nhiều ứng dụng; (ii) gắn với nhiều nền tảng lí thuyết. Liên quan đến công việc khai phá dữ liệu là *cách thức* thực hiện việc khai phá. Tuy hiện có nhiều tài liệu, nhưng một số tài liệu không đề cập những khía cạnh nền tảng về lí thuyết, kiến thức liên quan, mà chỉ nêu các đặc trưng và thuật toán cho các công cụ khai phá. Do vậy mà:

- Điều đầu tiên cần xem xét là lựa chọn công cụ khai phá khi xây dựng năng lực khai phá dữ liệu kinh doanh trong công ty. Có thể tìm được công cụ phù hợp với điều kiện của doanh nghiệp trên các trang mạng;
- Khía cạnh cần xem xét nữa là (i) tạo dựng đội công tác khai phá dữ liệu; (ii) xây dựng nền cho khai phá dữ liệu; (iii) hiểu biết kiến thức nền đối với thực hành khai phá dữ liệu;
- Điều quan trọng khác là cơ sở cho những điều cần thực hiện. Đó là năng lực về phương pháp luận khai phá và thống kê, có được qua kinh nghiệm làm việc. Phương pháp luận phân tích hình thành trong ngữ cảnh của lí thuyết giải tích và thống kê. Tuy nhiên đa số các ứng dụng chỉ lặp lại phương pháp luận phân tích của y học; nên ngay từ đầu quá trình khai phá dữ liệu, cần phát triển phân tích thống kê và khai phá dữ liệu hiện đại.

Để hiểu những hạn chế và sức mạnh của phương pháp luận và sử dụng nó hiệu quả, cần xét lí thuyết thống kê. Lí thuyết này không chỉ đề cập cách giải bài toán phân tích, mà cần thiết để lựa chọn cách thức thực hiện khác nhau. Các kĩ thuật thống kê phải dựa trên lí thuyết thống kê. Bởi vậy việc nghiên cứu về thống kê và lịch sử của thống kê là cần thiết. Các phương pháp khai phá dữ liệu đều dựa vào thống kê, nên thống kê ngày càng cần thiết đối với các nhiệm vụ trong các dự án về khai phá dữ liệu.

1.1.1. Lịch sử của khai phá dữ liệu và thống kê

*Phân tích mẫu*¹ là công việc từ xa xưa, nhằm xử lí dữ liệu và hiểu biết về dữ liệu. Người ta đã có khái niệm về trung bình và nhóm từ thế kỉ VI trước Công nguyên tại Trung Hoa cổ đại, ngay sau khi có bàn tính gậy. Bàn tính gậy là công cụ tính toán xuất hiện ở Trung Hoa vào khoảng năm 300 trước Công nguyên. Các nhà thống kê ở Trung Hoa cổ và Hy Lạp đã sử dụng thống kê trong những công việc liên quan đến quân sự và tài chính.



Hình 1.2. Bàn tính gậy xuất hiện tại Trung Hoa

Vào thế kỉ XVI, XVII, các trò chơi cần đến lựa chọn may rủi trở nên phổ biến; chúng liên quan đến xác suất. Một số nhà toán học nghiên cứu về vấn đề này là Fermat, Leibnitz... Các chủ đề nghiên cứu của họ dẫn đến các nghiên cứu về toán học và thống kê trong các năm về sau.

¹ Pattern: khuôn mẫu; sample: mẫu; template: hình mẫu.



Hình 1.3. Nhà toán học G. W. Leibnitz (1646 - 1716)

1.1.2. Thống kê hiện đại

Hai ngành phân tích thống kê phát triển trong thế kỉ XVIII là (i) thống kê Bayes; (ii) thống kê cổ điển.

Đối với thống kê Bayes, xác suất xảy ra sự kiện là bằng xác suất xảy ra quá khứ nhân với khả năng xảy ra trong tương lai. Phân tích dựa trên khái niệm về xác suất điều kiện có nghĩa xác suất của sự kiện xảy ra được biết qua sự kiện khác, đã xảy ra. Phân tích Bayes bắt đầu bằng việc tìm tri thức đã biết, tin cậy và giả thiết. Người ta đưa ra khái niệm *tiên nghiệm chủ quan*¹, kết hợp với dữ liệu quan sát ngẫu nhiên, nhờ hàm khách quan.



Hình 1.4. Nhà toán học T. Bayes (1702 - 1761)

Tiếp cận thống kê hiện đại, tiếp theo công trình toán của Gauss và Laplace, xét *xác suất liên kết*², thay vì xác suất điều kiện.

¹ Subjective prior: tiên nghiệm chủ quan.

² Joint probability: xác suất liên kết.

Đây là cơ sở phù hợp với công tác phân tích. Hàm xác suất liên kết diễn tả xác suất X nhận giá trị x đồng thời với Y nhận giá trị y , như hàm của x và y .

Lợi ích của xác suất là rõ ràng trong sinh học Mendel, cuối thế kỉ XIX. Nhiều học giả đóng góp cho ngành xác suất như Francis Galton; Karl Pearson phát triển khái niệm về hồi quy và tương quan trong phân tích dữ liệu gen; R. A. Fisher phát triển hệ thống thử suy luận trong nghiên cứu y khoa, dựa trên khái niệm độ lệch chuẩn. Các tiếp cận Bayes dùng trong quá trình suy luận cũng cho kết quả rộng trong nhiều ứng dụng y khoa, do chúng sử dụng các tập khác nhau về chủ quan tiên nghiệm. Đích của Fisher là phát triển hệ thống suy luận thống kê cho y khoa, với tập các công cụ so sánh hiệu quả trị liệu. Fisher đã đưa ra giả thiết về mô hình tham số.



Hình 1.5. Nhà thống kê R. Fisher

Sau đây là một số giả thiết đối với mô hình thống kê tham số.

1.1.2.1. Dữ liệu phù hợp với phân phối đã biết

Tác phẩm đầu tiên của Fisher dựa trên tính toán *độ lệch chuẩn* của tham số, với giả thiết rằng dữ liệu được phân phối theo phân phối *chuẩn*. Phân phối chuẩn có hình chuông, với trung bình tại đỉnh chuông. Độ lệch chuẩn đơn giản là trung bình của độ lệch tuyệt đối của các giá trị so với giá trị trung bình. Phát triển mô hình này bằng cách dùng tham số thay đổi, dựa trên *phân phối cung ứng*¹ hay phân phối Poisson.

¹ Logistic: cung ứng.

Giả thiết của phân phối cụ thể đã biết là cần thiết, để biết các đặc trưng của hàm phân phối, khi suy luận. Tất cả những phương pháp tham số đều theo tiếp cận khớp dữ liệu của thế giới thực với mô hình toán học.

1.1.2.2. Độc lập nhân tố

Trong các hệ thống dự đoán có tham số, biến sẽ dự đoán Y được xem là hàm của biến X . Giả thiết các biến X tác động độc lập lên biến Y . Tác động lên Y của mỗi biến X là không phụ thuộc vào tác động của biến X khác lên Y . Tình huống này giống như được tạo trong phòng thí nghiệm, nhờ chỉ dùng một biến thay đổi, trong khi giữ các biến, nhân tố khác cố định.

Tuy nhiên, hoàn cảnh thực tế khác với điều kiện trong phòng thí nghiệm. Một vài nhân tố, có tác động đến nhân tố khác, được phép tác động liên kết lên Y . Vấn đề này được gọi là *cộng tuyến*¹. Khi điều này xảy ra giữa nhiều nhân tố, người ta có khái niệm *đa cộng*². Vấn đề đa cộng đưa các nhà thống kê sử dụng thuật ngữ tương tác trong mối quan hệ, nhằm thể hiện tác động kết hợp.

1.1.2.3. Cộng tuyến tính

Người ta không chỉ muốn các biến X độc lập, mà còn muốn tác động của các biến X lên Y cần được tích lũy và là tuyến tính. Điều này có nghĩa mỗi nhân tố được cộng thêm hay bớt đi khỏi tác động kết hợp của tất cả các X lên Y . Nếu mối quan hệ giữa Y và các X không tuyến tính, người ta có thể sử dụng tương quan dạng hàm mũ. Giả thiết về cộng tuyến tính đối với mối quan hệ của các biến X có thể gây lỗi lớn lên đầu ra. Đây là trường hợp trong các hệ thống dữ liệu kinh doanh.

¹ Collinearity: cộng tuyến.

² Multicollinearity: đa cộng.

1.1.2.4. Biến hằng

Người ta còn có giả thiết liên quan đến các biến X là: sự thay đổi của các biến X trên miền giá trị của nó đều như nhau. Điều này có nghĩa nếu người ta chia dải của biến thành đoạn nhỏ, thì sai khác giữa các bản ghi đối với đoạn thứ nhất sẽ như đối với các đoạn khác trên dải của biến khác.

Nếu sai khác trên dải của biến là thay đổi, người ta nói rằng *biến ngẫu nhiên*¹. Lỗi trong giá trị dự đoán do biến ngẫu nhiên trong số các biến là lớn.

1.1.2.5. Các biến số và liên tục

Việc giả thiết rằng các biến cần là số và liên tục có nghĩa người ta giả thiết các dữ liệu có dạng số, hay có thể được chuyển sang dạng số trước khi đưa vào phân tích, và số này có phân phối liên tục.

Khi các giá trị trong mẫu dữ liệu là không liên tục, người ta nói chúng là *rời rạc*. Các phương pháp thống kê tham số cổ điển không kiểm định với dữ liệu rời rạc, do phân phối xác suất đối với dữ liệu liên tục khác với dữ liệu rời rạc. Thực tế trong cả khoa học lẫn kinh doanh, có nhiều bài toán với dữ liệu rời rạc, nhưng người ta vẫn sử dụng các phương pháp thống kê cổ điển với dữ liệu liên tục.

Như vậy có sự phân biệt ứng với hai tiếp cận trong thống kê, ứng với *sự liên tục* hay *sự rời rạc*, ứng với *tính toàn thể* hay *tính cục bộ*. Trong suy nghĩ của mỗi con người cũng thường có hai tiếp cận như vậy. Hai tiếp cận, hay hai khuynh hướng này đã là chủ đề tranh luận giữa những nhà triết học cổ đại Plato và Aristotle.

¹ Heteroscedastic: biến ngẫu nhiên.



Hình 1.6. Aristotle (384 - 322 trước Công nguyên)
và Plato (428 - 348 trước Công nguyên)

1.1.3. Hai góc nhìn về thực tại

Cho dù người ta tiến hành giải vấn đề hay trả lời câu hỏi cụ thể, cần bắt đầu với các khái niệm. Thường thì người ta cần thực hiện một trong hai cách (i) cố gắng giảm phức tạp bằng cách đưa về các phần tử chính; hay (ii) nhận thức bài toán theo thuật ngữ tổng quát. Đó là hai tiếp cận (i) chi tiết; hoặc (ii) toàn thể. Điều phân biệt này có xuất xứ từ nền triết học Hy Lạp, mà các nhà triết học Aristotle và Plato là đại diện.

1.1.3.1. Nhìn nhận theo nhà triết học Aristotle

Nhà triết học Aristotle tin rằng chân lí, sự thật về vạn vật có được chỉ do mắt nhìn thấy, tay sờ được... Ông tin rằng mức cao nhất của hoạt động trí tuệ là nghiên cứu chi tiết về thế giới hiện hữu xung quanh mình. Chỉ theo cách đó, người ta mới hiểu được sự thật.

Theo lập luận ấy, Aristotle tin rằng có thể chia hệ thống phức tạp thành từng mảnh. Tiếp theo là mô tả chi tiết các mảnh đó, rồi ghép các mảnh lại để hiểu *toàn bộ*. Đối với ông, toàn bộ được coi là tổng hòa của các thành phần. Bản chất của toàn bộ được ông nhìn nhận một cách máy móc.

Khoa học sớm nghiêng về lập luận của Aristotle. Bản chất của thế giới xung quanh ta được nghiên cứu qua việc nhìn sâu vào các

thành phần vật lí và các đơn vị, loài, sinh học. Chẳng hạn, hiểu biết về hệ thống tự nhiên được phát triển từ khái niệm của hệ sinh thái. Hệ sinh thái là hệ thống tự cung tự cấp, có tác động lẫn nhau. Người ta thấy nhiều đặc tính của hệ sinh thái được giải thích được bằng tiếp cận truyền thống, tức của Aristotle.

Tuy nhiên tiếp cận này chưa hoàn thiện. Để thấy được khiếm khuyết trong quá trình lập luận của Aristotle, trong định nghĩa đầy đủ về thế giới, người ta xét tiếp cận của nhà triết học Plato.

1.1.3.2. Nhìn nhận theo nhà triết học Plato

Nhà triết học Plato được Aristotle dạy dỗ 20 năm; cả hai nhất trí về bản chất của tự nhiên. Khi Aristotle mô tả mọi vật hữu hình trong thế giới bằng nghiên cứu chi tiết, Plato tập trung nghiên cứu thế giới gồm các ý tưởng, đằng sau cái hữu hình.

Đối với Plato, thứ duy nhất bền vững là *tư tưởng*. Ông tin rằng điều quan trọng nhất về sự tồn tại của loài người là thứ nằm đằng sau những thứ mà mắt con người trông thấy, tức những thứ ngoài tầm tay của con người cụ thể. Ông tin rằng ảnh hưởng của tư tưởng vượt lên trên thế giới hữu hình, tức vượt lên trên những thứ mà Aristotle quan tâm. Với Plato, toàn bộ của hiện thực là *lớn hơn* tổng của các thứ hữu hình.

Nền khoa học tại các nước phương Tây đã xây dựng các khái niệm về bản chất tự nhiên. Các nhà khoa học phương Tây đã theo nền tảng của triết học Plato. Triết học Plato có trên 2000 năm, đến thời *Ánh sáng*.

Nhưng về sau người ta cũng quay về với triết học của Aristotle. Việc phân nhánh các tiếp cận nghiên cứu về thế giới, liên quan đến tư tưởng và bản chất của thực tại thể hiện trong nhiều cuộc tranh luận giữa những nhà khoa học. Cũng qua các tranh luận ấy, mà người ta nảy sinh ra so sánh (i) chi tiết so với toàn thể; (ii) trên xuống so với dưới lên; (iii) não trái so với não phải...

1.1.4. Phân tích thống kê hiện đại: thể hệ hai

1.1.4.1. Phân tích thống kê với dữ liệu phi tuyến

Vào năm 1980, các nhà nghiên cứu về toán thống kê thấy rõ rằng hạn chế của tiếp cận Aristotle trong phân tích mối quan hệ phi tuyến trong bộ dữ liệu lớn của hệ thống phức tạp trong thế giới thực. Nghiên cứu toán học tiếp tục phát triển theo tiếp cận của nhà thống kê Fisher, bằng cách phát triển các thể hệ phi tuyến để xử lý tính phi tuyến trong phân phối dữ liệu liên tục. Tuy nhiên nhiều bài toán phi tuyến yêu cầu phân phối rời rạc hơn là liên tục. Các phương pháp này gồm:

- *Mô hình logit¹, gồm cả hồi quy cung ứng*. Giả sử dữ liệu theo phân phối cung ứng và biến phụ thuộc có kiểu *phạm trù*², tức 1/0. Trong mô hình này, biến Y phụ thuộc được xác định như hàm mũ của biến X . Mối quan hệ này có thể được xem xét đối với tính phi tuyến giữa Y và các X , chứ không đối với tương tác giữa các biến X ;
- *Mô hình probit³, gồm cả hồi quy Poisson*. Mô hình này giống mô hình logit, nhưng dữ liệu có phân phối Poisson, thay vì phân phối cung ứng;
- *Mô hình tuyến tính tổng quát GLM⁴*. Mô hình GLM mở rộng phương trình ước lượng dùng để dự đoán, $Y = f(X)$, f là hàm, X là vectơ biến dùng để dự đoán. Vế trái đẳng thức là *thành phần tất định*, vế phải là thành phần ngẫu nhiên. Các nhà thống kê ghi nhận (i) thành phần tất định có thể được mô tả như hàm mũ, giống hàm cung ứng; (ii) thành phần ngẫu nhiên tích lũy tác động của các biến X và vẫn là tuyến tính; và (iii) hàm liên kết có thể so sánh logic. Dấu bằng được gọi là *liên kết đồng nhất*⁵. Hiện tại các nhà toán

¹ Hàm logit là hàm ngược của hàm cung ứng.

² Category: phạm trù.

³ Probit: đơn vị xác suất dựa trên độ lệch giữa trung bình của phân phối chuẩn.

⁴ GML: Generalized Linear Model: mô hình tuyến tính tổng quát.

⁵ Identity link: liên kết đồng nhất.

học đã có khung để xác định hàm phù hợp với bộ dữ liệu phi tuyến mạnh. Mạng nơron cho phép diễn tả các hàm với mức phi tuyến bất kì.

Trong khi người ta sử dụng nhiều phương pháp thống kê theo tiếp cận Fisher, có nhóm nghiên cứu phát triển tiếp cận của nhà toán học Bayes.

Theo tiếp cận Bayes, sự kiện xảy ra trong quá khứ, tức các dấu hiệu thực tế, là có ý nghĩa hơn dấu hiệu thống kê tính được nhờ các hàm xác suất liên kết. Chẳng hạn trị liệu cần chẩn đoán chính xác ung thư, ứng với *đúng*¹, là quan trọng hơn lỗi do chẩn đoán nhầm, ứng với *sai*. Mở rộng tiếp cận này, người ta không theo tiếp cận Plato; tiếp cận Plato liên hệ chẩn đoán chính xác với môi trường dữ liệu, mà bất kì mẫu cụ thể nào đã vạch ra, chứ không với chính dữ liệu của riêng mẫu. Để trị liệu, người ta cần bỏ qua thực tế rằng chỉ xét xác suất của các sự kiện đã xảy ra trong quá khứ của môi trường dữ liệu, chứ không xét xác suất của các sự kiện *cần* xảy ra, nhưng đã không xảy ra. Liên quan đến điều này, có khái niệm về sai số alpha và sai số beta.

Định nghĩa: Sai số alpha là xác suất là sai khi người ta nghĩ mình đúng.

Định nghĩa: Sai số beta là xác suất là đúng khi nghĩ mình sai.

Trong thống kê Fisher, *quan sát* và *sai số alpha* tương ứng xác định sự khác biệt giữa sự kiện xảy ra với sự kiện mong đợi. Tiếp cận Fisher thiết lập sai số alpha ngay từ đầu phân tích và tham chiếu đến sai số khác có ý nghĩa giữa những *dân cư*² dữ liệu trong khuôn khổ sai số alpha được xác định. Người ta cũng thêm hậu tố vào dự đoán, chỉ độ tin cậy. Độ tin cậy là phần bù của sai số alpha; tức tin cậy 95% và sai số alpha 5%. Sai số beta được dùng để chỉ độ mạnh, hay độ vững vàng, của thử nghiệm phân tích.

¹ Các khái niệm *true positive*, *false positive*, *true negative* và *false negative* gắn với đánh giá độ chính xác của tìm kiếm.

² Population: dân cư.

Tiếp cận Bayes không xử lý các sai số như tiếp cận Fisher. Họ lập luận rằng không có quyết định đúng nếu không xét các phương án thay thế một cách cẩn thận. Họ cho rằng mức xác suất tính được 0.02 đối với sự kiện đã cho trong dữ liệu mẫu không ngầm xác định xác suất 0.02 về sự kiện này trong thế giới các sự kiện.

Tiếp cận nào đúng phụ thuộc vào (i) bản chất của nghiên cứu, (ii) khả năng của xem xét tiên nghiệm và (iii) chi phí liên quan của các lỗi dương sai và âm sai. Trước khi lựa chọn, người ta thấy tất cả các phép thử thống kê đều có ưu, nhược điểm. Cần dựa vào độ mạnh, yếu của tiếp cận, để hiểu rõ về kết quả thu được. Trong kinh doanh, người ta chú trọng các thay thế *khả thi*.

1.1.4.2. Phân tích thống kê với dữ liệu ở mọi nơi

Nhu cầu thực tế trong kinh doanh về *trích rút*¹ tri thức từ dữ liệu tăng nhanh. Điều này cần đến các kĩ thuật phân tích mới, cho phép phân tích mối quan hệ phi tuyến trong bộ dữ liệu lớn, với phân phối chưa rõ.

Việc phát triển kĩ thuật mới đòi hỏi ba hướng tiếp cận, chứ không phải hai hướng theo Bayes hay Fischer như vừa nêu. Tiếp cận thứ ba, liên quan đến học máy, có thể xem như dựa vào các tiếp cận về chân lí của Aristotle và Plato, không dựa vào tiếp cận Bayes.

1.1.5. Các phương pháp học máy: thể hệ ba

Nghiên cứu về học máy xuất phát từ trí tuệ nhân tạo, khi để cập máy thông minh. Ban đầu các phương pháp này theo hai hướng song song (i) mạng nơron; (ii) cây quyết định.

1.1.5.1. Học máy theo mạng nơron nhân tạo

Mạng nơron nhận tạo ANN² giải quyết trực tiếp vấn đề về hàm phi tuyến, với nghĩa gán trọng số cho các biến đầu vào, thu thập tác

¹ Extract: trích, trích rút.

² ANN: Artificial Neural Network: mạng nơron nhân tạo.

động của chúng, để sinh được đầu ra, theo loại hàm quyết định. Các hệ thống này thể hiện hệ thống thần kinh của người. Việc di chuyển các xung nhanh hay chậm trong não người là biến số.

Mối quan hệ phức tạp của các nơron và liên kết với khớp thần kinh được huấn luyện, và sẽ học để đáp ứng nhanh hơn. Các nhà khoa học máy tính bắt đầu diễn tả loại hệ thống này dưới dạng mạng nơron nhân tạo, cho phép học cách ghi nhận các *khuôn mẫu*¹ trong các biến đầu vào của bộ dữ liệu.

1.1.5.2. Học máy theo cây quyết định

Phát triển học máy thứ hai liên quan đến diễn tả *tác động* trực tiếp của các phương pháp đang phát triển, để phát hiện các *luật* mà có thể được đánh giá đối với tách các giá trị đầu vào thành một hay nhiều khối, mà không phải mô tả mối quan hệ phiếm hàm một cách trực tiếp.

Các phương pháp này tập trung diễn tả các luật hiện, tức luật quy nạp, hay diễn tả mối quan hệ giữa các luật, tức cây quan hệ; các luật này biểu hiện kết quả. Các phương pháp này tránh sự chặt chẽ của mô hình tham số, và phù hợp để phân tích các sự kiện phi tuyến NLE², cả theo cách tác động kết hợp của các biến X lên Y và tương tác giữa các biến độc lập.

Trong khi các cây quyết định và các mạng nơron có thể diễn tả các NLE đầy đủ hơn các phương pháp thống kê tham số, chúng vẫn là tuyến tính trong các hàm gộp của chúng.

1.1.6. Lí thuyết học thống kê: thể hệ thứ tư

Các kĩ thuật hồi quy cung ứng có thể tính tác động *kết hợp* của các tương tác, trong số tất cả các biến X, với tư cách hàm phi tuyến xác định biến phụ thuộc Y. Vẫn còn những hạn chế đối với máy

¹ Pattern: khuôn mẫu (để tránh nhầm với mẫu *sample*).

² NLE: Nonlinear Event: sự kiện phi tuyến.

học tuyến tính. Cho dù mạng nơron và cây quyết định đã giải quyết vấn đề này, người ta thấy vẫn cần mở rộng việc học máy. Một cách diễn tả các hạn chế này là thể hiện việc học theo *không gian giả thuyết*.

Định nghĩa: Không gian giả thuyết là kiến trúc mà trong đó người ta tìm thấy giải pháp.

Nhưng không gian của các lời giải chấp nhận được có thể bị ràng buộc chặt bởi các hàm tuyến tính trong phân tích thống kê cổ điển và các kĩ thuật học máy. Các vấn đề phức tạp trong thế giới thực sẽ cần không gian giả thuyết có sức diễn tả hơn, để đáp ứng ràng buộc phi tuyến. Các mạng nơron nhiều tầng cho phép tính nhiều tác động phi tuyến với danh nghĩa kiến trúc mạng và kĩ thuật tối thiểu sai số.

Tiếp cận khác là thể hiện các điểm dữ liệu như các vector, giống như các dòng bản ghi. Các vector ấy gồm các thành phần, như các thuộc tính của bản ghi. Không gian vector của các dòng được đặc trưng theo cách khái niệm và toán học, như không gian N chiều; trong đó N là số thuộc tính. Khi hiện dữ liệu về bản ghi như vector, người ta sử dụng điểm mạnh của đại số tuyến tính; có thể diễn tả tất cả sai khác giữa các thuộc tính của hai bản ghi, bằng cách tính *tích chấm*¹, hay *tích nội*². Tích chấm của hai vector là tổng của tích các thuộc tính tương ứng. Sau đó người ta có thể diễn tả dữ liệu như dãy các tích chấm, vào không gian tích nội N chiều.

Các thuật toán thống kê cổ điển, như hồi quy tuyến tính, có thể được diễn tả theo cách này. Trong lí thuyết học thống kê, người ta dùng nhiều hàm phức, hay *nhân*³, thay thế tích nội. Khi khớp dữ liệu vào các không gian nhân phức, phạm vi của các lời giải đối với một bài toán tăng đáng kể. Dữ liệu trong không gian

¹ Dot product: tích chấm.

² Inner product: tích nội.

³ Kernel: nhân.

này được tham chiếu là *đặc trưng*, thay vì các thuộc tính trong dữ liệu gốc.

Một số kĩ thuật học mới có ưu điểm về thuộc tính của máy học nhân. Thực hiện chung nhất là máy vectơ tựa. Khi mạng nơron được huấn luyện, các dòng dữ liệu đưa vào mạng, và tính được sai số giữa các giá trị quan sát và dự đoán, như học có giám sát. Hàm học cho huấn luyện và hàm tối thiểu sai số gắn nhau chặt chẽ trong mạng nơron. Nhưng không như vậy đối với máy vectơ tựa. Do quá trình học tách rời quá trình xấp xỉ, người ta có thể thực nghiệm nhờ dùng các xác định nhân khác, với lí thuyết học khác. Do vậy, thay cho việc chọn trong số các kiến trúc khác nhau cho ứng dụng mạng nơron, người ta có thể thực nghiệm với các nhân khác trong máy vectơ tựa.

Các phương pháp thể hệ thứ tư chưa kết thúc. Khi tích lũy được nhiều dữ liệu, người ta có thể phát minh ra cách thông minh hơn để mô phỏng quá trình học của con người.

Chiến lược mới nhằm vào liên kết các máy tính, như nhiều nơron của não người, với (i) tính toán lưới; (ii) tính toán đám mây.

Định nghĩa: Tính toán lưới là tính toán dùng nhóm các máy tính nối mạng, để chia và xử lí các vấn đề tính toán.

Các chiến lược này sử dụng nhiều máy tính, với hi vọng thuận lợi hơn khi khai phá dữ liệu. Tiếp cận phân tích với nhiều máy tính là bước logic tiếp theo của việc phát triển não người. Bước này có thể phát triển thành thể hệ năm của khai phá dữ liệu.

Định nghĩa: Tính toán đám mây là tính toán dùng Internet để phân phối dữ liệu và các nhiệm vụ tính toán đến nhiều máy tính mọi nơi trên thế giới, nhưng không cần hạ tầng phần cứng tập trung như tính toán lưới.

1.2. Khía cạnh lí thuyết đối với khai phá dữ liệu

Việc phân tích thống kê đã có từ cổ đại, nhưng việc khai phá dữ liệu chắc chắn mới có ở thời hiện đại, tức gần đây. Nó phát triển vào những năm 90 của thế kỉ trước, và hợp nhất thành một

lĩnh vực trong đầu thế kỉ XXI. Khai phá dữ liệu thể hiện kết hợp của vài lĩnh vực như (i) phân tích thống kê truyền thống; (ii) trí tuệ nhân tạo; (iii) học máy; và (iv) phát triển cơ sở dữ liệu lớn.

Phân tích thống kê truyền thống tiếp tục *phương pháp suy diễn* khi tìm kiếm mối quan hệ trong bộ dữ liệu. Trí tuệ nhân tạo, như hệ chuyên gia, và các kĩ thuật học máy tiếp tục *phương pháp quy nạp* để phát hiện các khuôn mẫu về mối quan hệ trong bộ dữ liệu. Lập luận suy diễn là quá trình như Aristotle về phân tích dữ liệu chi tiết, tính số đo, và tạo nên kết luận dựa trên kiến thức toán học về độ đo. Quy nạp là quá trình như của Plato về sử dụng thông tin trong bộ dữ liệu, tạo nên kết luận tổng quát, dù không hoàn toàn trực tiếp chứa các dữ liệu đầu vào. Phương pháp khoa học theo tiếp cận quy nạp, nhưng có nhiều yếu tố như tiếp cận Aristotle trong những bước thực hiện đầu.

1.2.1. Phương pháp khoa học

Phương pháp khoa học theo quy trình chín bước sau:

1. Xác định vấn đề;
2. Thu thập thông tin đã có về hiện tượng;
3. Tạo một hay nhiều giả thuyết;
4. Thu thập dữ liệu thực nghiệm mới;
5. Phân tích thông tin trong bộ dữ liệu mới;
6. Diễn giải kết quả;
7. Tổng hợp kết luận, dựa trên dữ liệu cũ, dữ liệu mới và *trực giác*;
8. Tạo giả thuyết mới cho lần thử nghiệm sau;
9. Lặp lại quá trình.

Bước 1 - 5 yêu cầu suy diễn; bước 6 - 9 yêu cầu quy nạp. Dù phương pháp khoa học dựa nhiều vào lập luận suy diễn, sản phẩm cuối cùng là do quy nạp, suy diễn. Khai phá dữ liệu giống như vậy. Thực tế, các thuật toán học máy dùng trong khai phá dữ

liệu được thiết kế để bắt chước quá trình xây trong suy nghĩ của các nhà khoa học, như người ta hay nói là *đầu* khoa học.

Khai phá dữ liệu sử dụng toán học, những kết quả *không* xác định một cách như toán học. Đó là yếu tố cảm tính con người trong kết quả khai phá.

1.2.2. Một số khái niệm liên quan đến khai phá dữ liệu

Khai phá dữ liệu được định nghĩa theo nhiều cách khác nhau, do nhìn nhận khác nhau về các khía cạnh của khai phá dữ liệu. Một định nghĩa của Frawley, năm 1991, là:

Định nghĩa: Khai phá dữ liệu là trích ra điều không tầm thường từ những thứ chưa biết trước, ngầm định, và trích ra thông tin hữu dụng tiềm tàng từ dữ liệu.

Do khai phá dữ liệu phát triển như hoạt động chuyên nghiệp, người ta phân biệt nó với hoạt động trước đây và mô hình hóa thống kê và các hoạt động rộng về phát hiện tri thức. Một số khái niệm liên quan được xác định:

Định nghĩa: Mô hình hóa thống kê là việc sử dụng các thuật toán thống kê có tham số để nhóm hay dự đoán đầu ra hay sự kiện, dựa trên các biến đầu vào.

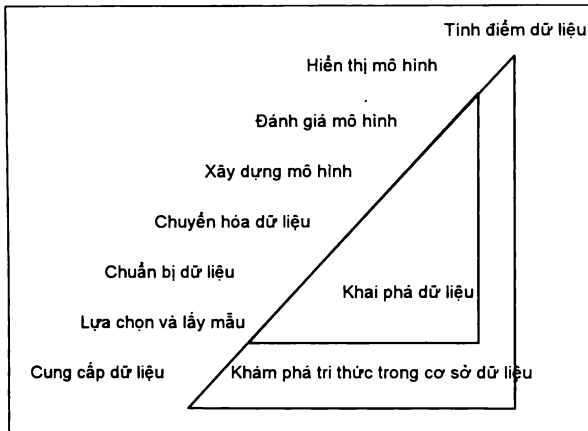
Xác định lại khái niệm về khai phá dữ liệu:

Định nghĩa: Khai phá dữ liệu là quá trình sử dụng các thuật toán học máy để phát hiện các khuôn mẫu mờ nhạt về mối quan hệ giữa các thành phần dữ liệu trong bộ dữ liệu lớn, lộn xộn, có nhiễu, để dẫn đến hành động tăng lợi ích dưới dạng nào đó, chẳng hạn như chẩn đoán, phát hiện lỗi,...

Khám phá tri thức¹ là lĩnh vực rộng hơn, đi song song với khai phá dữ liệu:

¹ Knowledge discovery: khám phá tri thức. Thuật ngữ này còn là Knowledge Exploration. Đề nghị sử dụng thuật ngữ khám phá, để không trùng với thuật ngữ khai phá, Minning.

Định nghĩa: Khám phá tri thức là toàn bộ quá trình truy cập dữ liệu, thăm dò dữ liệu, chuẩn bị dữ liệu, mô hình hóa, sử dụng mô hình và giám sát mô hình.



Hình 1.7. Quan hệ giữa khai phá dữ liệu và khám phá tri thức

Quá trình gồm khai phá dữ liệu được minh họa trong hình trên. Các định nghĩa chú trọng đến khía cạnh đặc biệt của thông tin và nguồn thông tin. Năm 1996, U. Fayyad đã đề xuất: *Khám phá tri thức trong các cơ sở dữ liệu là quá trình không tầm thường, để định danh các khuôn mẫu được kiểm chứng, mới, sử dụng tiềm năng và hiệu được, trong dữ liệu.*

Định nghĩa thứ hai về khai phá dữ liệu tập trung vào *khuôn mẫu* trong dữ liệu, thay vì thông tin theo nghĩa tổng quát. Các khuôn mẫu này là mờ nhạt và khó phân biệt, và chỉ được *cảm nhận* nhờ thuật toán phân tích; thuật toán này ước lượng mối quan hệ phi tuyến giữa các biến đầu vào X và đích Y , và giữa các X . Định nghĩa này phát triển theo các công cụ của học máy trong khai phá dữ liệu. Các công cụ như cây quyết định và mạng nơron cho phép phân tích các khuôn mẫu phi tuyến trong dữ liệu thuận tiện hơn so với các thuật toán thống kê có tham số. Lí do ở chỗ các thuật toán học máy học theo cách con người, chẳng hạn không tính độ đo dựa trên trung bình và phân phối dữ liệu.

Định nghĩa về khai phá dữ liệu đã hạn chế về quá trình xây dựng mô hình. Nhưng trong thực tế, các bộ công cụ khai phá dữ liệu, như SPSS¹ - Clementine, có cả công cụ khác để thuận tiện cho xây dựng mô hình và cho đánh giá và hiển thị mô hình.

Khám phá tri thức trong cơ sở dữ liệu KDD² kết hợp những công cụ toán học, để phát hiện các khuôn mẫu quan tâm, với toàn bộ quá trình trích dữ liệu và sử dụng các mô hình thu được, trên bộ dữ liệu. Khám phá tri thức làm tăng thông tin theo đích nào đó. Quá trình này pha trộn công nghệ về hệ thống kinh doanh, các phương pháp thống kê và năng lực tính toán, để phát hiện *cấu trúc*, tức các liên kết, khuôn mẫu, mối liên quan, chức năng cơ bản..., thay vì các tham số thống kê, tức trung bình, trọng số, ngưỡng...

1.2.3. Khung lý thuyết cho quá trình khai phá dữ liệu

Bản chất tiến hóa của định nghĩa và trọng tâm của khai phá dữ liệu là thực nghiệm. Vấn đề chính của phát triển này là chưa có lý thuyết đảm bảo bao trùm các khía cạnh về thông tin, nguồn thông tin, cách chuyển hóa thông tin... Khái niệm logic này được gọi là *lý thuyết mô hình*.

Định nghĩa: Lý thuyết mô hình là cái liên kết logic với các biểu thức đại số về cấu trúc, để mô tả hệ thống hay quá trình phức, với cú pháp bên vững và mối quan hệ giữa chúng, tức ngữ nghĩa.

Hầu hết các biểu hiện của các hoạt động khai phá dữ liệu có các thuật ngữ bên vững, chẳng hạn thuộc tính..., có thể ngầm cho biết quan hệ ngữ nghĩa logic khác với các phần tử dữ liệu sử dụng.

Năm 2000, Mannila tổng kết số tiêu chí cần thỏa mãn trong một tiếp cận, để phát triển lý thuyết mô hình cho khai phá dữ liệu. Các tiêu chí này có khả năng:

¹ SPSS: Statistical Package for the Social Sciences: bộ công cụ thống kê cho khoa học xã hội.

² KDD: Knowledge Discovery in Databases: khám phá tri thức trong cơ sở dữ liệu.

1. Mô hình hóa các nhiệm vụ khai phá dữ liệu điển hình, tức phân cụm, phát hiện luật, phân lớp;
2. Mô tả dữ liệu và tổng quát theo quy nạp, từ dữ liệu;
3. Biểu hiện thông tin từ nhiều dạng dữ liệu, tức dữ liệu quan hệ, chuỗi, văn bản...;
4. Trợ giúp tương tác và quá trình lặp;
5. Biểu hiện mối quan hệ toàn diện;
6. Kết hợp người dùng trong quá trình;
7. Kết hợp nhiều tiêu chí để xác định phương pháp khám phá quan tâm.

Người ta cũng mô tả các tiếp cận để phát triển lý thuyết mô hình khả dĩ, tuy chưa có tiếp cận nào đáp ứng tất cả tiêu chí trên.

1.2.3.1. Tiếp cận kinh tế vi mô

Bắt đầu của tiếp cận kinh tế vi mô là khai phá dữ liệu liên quan đến phát hiện các *khuôn mẫu* hoạt động được trong dữ liệu, mà có vài *tiện ích* để ra quyết định.

Đích của tiếp cận kinh tế vi mô là phát hiện quyết định, để tối đa lợi ích cho người dùng.

1.2.3.2. Tiếp cận cơ sở dữ liệu quy nạp

Cơ sở dữ liệu quy nạp gồm tất cả dữ liệu theo cấu trúc đã cho, cùng với tất cả những câu hỏi về các *khuôn mẫu* trong dữ liệu. Các sự kiện được lưu trữ và suy diễn đều được quản lý theo cùng một cách. Một trong những chức năng quan trọng của não người là phát hiện *khuôn mẫu*. Các dữ liệu chi tiết sẽ nhiều lên trong bộ nhớ, và các hành động được thực hiện theo *khuôn mẫu* đã lưu trữ.

Mannila cho rằng (i) tiếp cận kinh tế vi mô có thể biểu hiện hầu hết các yêu cầu đối với lý thuyết mô hình dựa trên các sự kiện đã lưu; nhưng (ii) tiếp cận cơ sở dữ liệu quy nạp biểu diễn dễ

dàng các sự kiện suy diễn được. Người ta phát triển ngôn ngữ đánh dấu mô hình hóa dự đoán PMML¹ như siêu tập của ngôn ngữ đánh dấu mở rộng XML².

1.2.4. Tính chắc chắn của quá trình khai phá dữ liệu

Thống kê truyền thống nghiên cứu việc sử dụng thông tin để *xác định* trạng thái tương lai của hệ thống, thường gọi là dự đoán, trong khi khai phá dữ liệu nghiên cứu việc sử dụng thông tin quá khứ để kiến thiết khuôn mẫu, dựa trên không chỉ dữ liệu vào, mà còn dựa vào *ảnh hưởng logic* của dữ liệu đó. Quá trình này cũng gọi là *dự đoán*, nhưng có phần tử mà không có trong phân tích thống kê: khả năng đảm bảo biểu diễn có thứ tự của *điều có thể* trong tương lai, so với *điều đã có* trong quá khứ dựa vào giả thiết của phương pháp thống kê.

So sánh với nghiên cứu phân tích thống kê, lĩnh vực khai phá dữ liệu phát hiện các khuôn mẫu và phân lớp, nhìn và dự đoán tương lai. Nhìn chung, khai phá dữ liệu cho phép (i) cung cấp hiểu biết toàn diện hơn về dữ liệu, nhờ phát hiện các khuôn mẫu mà trước đây chưa thấy; (ii) tạo mô hình để dự đoán, để ra quyết định, tạo dựng sự kiện tương lai.

1.2.5. Dữ liệu hướng người dùng so với dữ liệu hướng tính toán

Hầu hết các cơ sở dữ liệu trong kinh doanh được thiết kế với tiêu chí về hiệu quả lưu trữ và tìm kiếm hay sinh ra thông tin. Các phép kinh doanh được điều khiển nhờ hệ thống kế toán, tức ứng dụng công nghệ thông tin cho kinh doanh, theo cùng cấu trúc dữ liệu. Việc chú trọng vào cấu trúc dữ liệu trên các giao tác cũng thường xảy ra trong công việc kế toán. Dữ liệu trong các hệ thống

¹ PMML: Predictive Modeling Markup Language: ngôn ngữ đánh dấu mô hình hóa dự đoán.

² XML: Extended Markup Language: ngôn ngữ đánh dấu mở rộng.

kinh doanh theo giao tác ban đầu sẽ lưu trong cơ sở dữ liệu truy cập tuần tự theo chỉ số ISAM¹.

Khi dữ liệu tăng và cần sự linh động, người ta phát triển hệ quản trị cơ sở dữ liệu quan hệ RDBMS². Hệ quản trị này thuộc thế hệ hai.

Các hệ thống hướng tính toán là hoàn toàn hiệu quả cho mục đích định hướng, nhưng có nhược điểm: khó quản lí người dùng do việc tính toán. Một người có thể có một tài khoản hay nhiều tài khoản; một tài khoản có thể do nhiều người sở hữu. Như vậy RDBMS sẽ khó liên hệ công việc với người dùng xác định.

Khi lưu trữ dữ liệu theo dạng phù hợp với người dùng, việc trích thông tin để xây dựng bản ghi phân tích khách hàng CAR³ sẽ thuận tiện. Cơ sở dữ liệu hướng người dùng được xem là *cơ sở dữ liệu cho người dùng*.

1.2.5.1. Kho dữ liệu chuyên đề vật lí

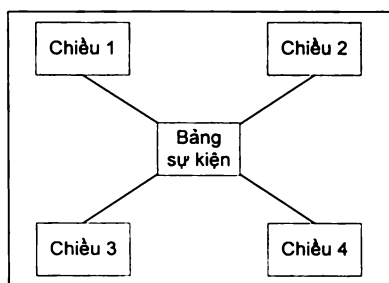
Một giải pháp cho vấn đề này là tổ chức cấu trúc dữ liệu để quản lí khía cạnh thông tin của khách hàng. Các cấu trúc này được thể hiện trên bảng, có khóa chung để nối các bảng. Kiến trúc cơ bản là dùng bảng sự kiện, nối với các bảng chiều; nó tạo nên *lược đồ sao*, hay cơ sở dữ liệu nhiều chiều.

Sẽ có cơ sở dữ liệu nhỏ, ứng với chiều quan tâm. Chúng phụ thuộc vào cơ sở dữ liệu chính, cho phép tạo các kết quả gộp lớn. Khi cơ sở dữ liệu chính là kho dữ liệu, thì các cơ sở dữ liệu nhỏ là cơ sở dữ liệu chuyên đề. Kho dữ liệu chuyên đề chứa dữ liệu trích ra từ kho dữ liệu.

¹ ISAM: Indexed Sequential Access Method: phương pháp truy cập tuần tự có chỉ số.

² RDBMS: Relational Database Management Systems: hệ quản trị cơ sở dữ liệu quan hệ.

³ CAR: Customer Analytical Record: bản ghi phân tích khách hàng.



Hình 1.8. Cơ sở dữ liệu lược đồ hình sao đơn giản

1.2.5.2. Kho dữ liệu chuyên đề ảo

Do năng lực tính toán và khả năng lưu trữ tăng, từ 1990 trở đi người ta đã quan tâm đến sử dụng đặc điểm và thông tin kế toán trước đây, dẫn đến việc quản trị quan hệ khách hàng CRM¹.

Thành công của thao tác RCM thông minh đã thay đổi cách công ty nhìn nhận về dữ liệu. Cơ sở dữ liệu không chỉ làm nhiệm vụ kế toán, mà còn quan sát khách hàng, với nghĩa mọi hoạt động kế toán, sản phẩm, dữ liệu địa lí... đều gắn với khách hàng. Tuy nhiên các cơ sở dữ liệu này thiết kế theo khung nhìn logic, gắn với một phần của cơ sở dữ liệu lớn. Nó tạo ra kho dữ liệu chuyên đề; do nó gắn với khung nhìn logic của RDBMS, kho này được xem là kho dữ liệu chuyên đề ảo.

1.2.5.3. Cơ sở dữ liệu cho gia đình

Một hình thức khác về dữ liệu hướng người dùng đã gắn tất cả những tính toán và các cá nhân vào cơ sở dữ liệu gia đình. Quá trình quản lí này đòi hỏi *khớp* tương đối để gộp tất cả những tính toán vào cùng một khách hàng. Khớp này sử dụng kiến thức về tập mờ, xem là *khớp mờ*.

Cấu trúc dữ liệu gia đình có thể có các bảng (i) kế toán; (ii) cá nhân; (iii) gia đình. Các dữ liệu lịch sử được kết hợp với mỗi mức

¹ CRM: Customer Relationship Management: quản trị quan hệ khách hàng.

phân cấp gộp lớn. Chẳng hạn bảng dữ liệu có thể hạn chế với dữ liệu hiện tại, hay dữ liệu quá khứ...

1.2.6. Thay đổi khung dữ liệu

1.2.6.1. Dữ liệu hướng người dùng

Tổ chức cấu trúc dữ liệu phù hợp đối với khai phá dữ liệu đòi hỏi thay đổi cơ bản cách nghĩ về dữ liệu trong kinh doanh. Dữ liệu không phục vụ tính toán, mà được tổ chức phục vụ khách hàng. Như vậy, dữ liệu cần tổ chức theo cấu trúc hướng người dùng, cho phép:

- Mọi quan hệ của tất cả phần tử dữ liệu cần hiện ra cho người dùng;
- Cấu trúc dữ liệu cần thuận tiện để chuyển sang dạng phù hợp với nhiệm vụ khai phá dữ liệu, chẳng hạn cho CAR.

1.2.6.2. Tạo bản ghi phân tích khách hàng CAR

Tất cả dữ liệu cần tải vào CAR. Quá trình này tương tự như khâu chuẩn bị. Nếu nguồn dữ liệu là kho dữ liệu, quá trình cần *giải chuẩn* dữ liệu, tức đưa dữ liệu về dạng phù hợp.

Định nghĩa: Giải chuẩn là quá trình trích dữ liệu từ các bảng theo chuẩn theo mô hình quan hệ của kho dữ liệu.

Dữ liệu trong các bảng liên kết với thuộc tính riêng. Việc trích dữ liệu sẽ đưa dữ liệu nguồn về bảng dữ liệu trong CAR. CAR sẽ gồm (i) tất cả phần tử dữ liệu đã tổ chức thành một bản ghi đối với mỗi khách hàng; (ii) một hay nhiều đích Y được suy diễn.

CAR thể hiện dạng văn bản như phương trình $Y = X_1 + \dots + X_n$. Kiến trúc dữ liệu được phân tích theo thống kê hay thuật toán học máy, theo phương pháp luận xác định. CAR được phân tích nhờ các thuật toán thống kê có tham số hay thuật toán học máy, trong quá trình KDD. Khía cạnh khai phá dữ liệu của KDD gồm một loạt các hoạt động nhằm huấn luyện và đánh giá các khuôn mẫu, hay

các phương trình tốt nhất. Các khuôn mẫu hay phương trình tối ưu được gọi là *mô hình*.

1.2.7. Các hoạt động chính của khai phá dữ liệu

Các hoạt động chính được liệt kê trong các mục nhỏ.

1.2.7.1. Khai phá dữ liệu như phân tích dữ liệu

Hoạt động khai phá dữ liệu theo cách phân tích dữ liệu bao gồm các kĩ thuật tương tác, đồ họa, cho phép người dùng thấy được bộ dữ liệu, theo các tham số thống kê.

Thể hiện đồ họa có ý nghĩa trực quan, tâm lí. Qua việc khai phá, người ta thấy được khuôn mẫu hay xu thế của bộ dữ liệu.

1.2.7.2. Mô hình hóa mô tả

Hoạt động mô hình hóa mô tả tạo khung nhìn mức cao về bộ dữ liệu. Các khung nhìn gồm:

1. Xác định phân phối xác suất tổng thể của dữ liệu, đôi khi gọi là *ước lượng mật độ*¹;
2. Các mô hình mô tả mối quan hệ giữa các biến, đôi khi gọi là *mô hình hóa phụ thuộc*;
3. Phân đoạn dữ liệu thành các nhóm, hoặc nhờ *phân tích cụm* hay *phân đoạn*². Phân tích cụm cố gắng phát hiện các nhóm tự nhiên, hoặc với nhiều cụm, hoặc một kiểu của phân tích cụm. Phân đoạn phát hiện các nhóm đồng nhất với biến mô hình hóa.

1.2.7.3. Mô hình hóa dự đoán theo phân lớp và hồi quy

Đích của mô hình hóa dự đoán là xây dựng mô hình mà giá trị của một biến có thể được dự đoán nhờ các giá trị của các biến

¹ Density estimation: ước lượng mật độ.

² Segmentation: phân đoạn.

khác. (i) Phân lớp dùng cho các biến theo dạng phạm trù, tức biến dạng có/không đa lựa chọn. (ii) Hồi quy dùng cho biến liên tục, tức các biến có giá trị số...

1.2.7.4. Khai phá khuôn mẫu và luật

Hoạt động này đòi hỏi bất kì điều gì để phát hiện tổ hợp các hạng mục thường xảy ra trong cơ sở dữ liệu giao tác. Phát triển của luật kết hợp được trợ giúp của nhiều thuật toán trong các sản phẩm phần mềm thương mại.

Một phương pháp liên quan là phân tích chuỗi, mối quan hệ và liên kết SAL¹. Phân tích SAL phát triển không chỉ mối quan hệ, mà còn phát triển chuỗi hạng mục có quan hệ. Từ các mối quan hệ xâu chuỗi này, người ta tính được các *liên kết*.

1.2.7.5. Tìm kiếm theo nội dung

Với khuôn mẫu đã biết và có đích, người ta tìm kiếm, phát hiện các khuôn mẫu phù hợp trong bộ dữ liệu mới. Tiếp cận này đối với nhận dạng mẫu thường được sử dụng với tư liệu văn bản hay bộ dữ liệu ảnh.

Định nghĩa: Tìm kiếm thông tin IR² là hoạt động thu thập các nguồn thông tin liên quan đến cái cần tìm kiếm, có thể dựa trên siêu dữ liệu hoặc trên việc đánh chỉ mục toàn văn, hoặc dựa trên nội dung khác.

Trái với hình ảnh về người khai phá dữ liệu như nhà ảo thuật, các hoạt động trên có nguyên lí đơn giản. Chúng thực hiện các hoạt động theo cách học của não người. Các thuật toán học máy học theo tình huống, như con người vậy. Dữ liệu nhập vào trí nhớ không theo cách riêng lẻ, mà theo *khuôn mẫu*. Các khuôn mẫu này so sánh với tập các tín hiệu thần kinh; thần kinh kết nối dữ liệu nhập với thông tin quá khứ. Não chính là máy nhận dạng mẫu lí

¹ SAL: Sequence, Association and Link: chuỗi, mối quan hệ và liên kết.

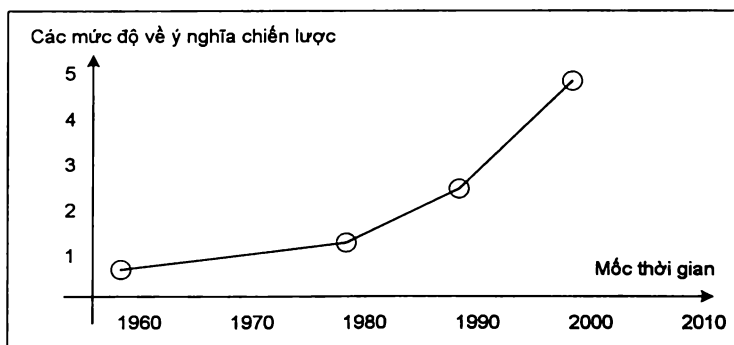
² IR: Information Retrieval: tìm kiếm thông tin.

tương. Các khuôn mẫu được ghi lại dưới dạng các trọng số, hằng số toán học hay các nhóm.

Khai phá dữ liệu không phải chuyên ngành hàn lâm mới, mà là nấc logic tiếp theo về phát triển trong kinh doanh, để sử dụng dữ liệu. Bảng sau cho thấy mốc lịch sử của khai phá dữ liệu.

Bảng 1.1. Mốc phát triển của khai phá dữ liệu

Bước phát triển	Thu thập dữ liệu	Truy cập dữ liệu	Kho dữ liệu và trợ giúp quyết định	Khai phá dữ liệu
Vấn đề thương mại	Thu nhập trong năm qua?	Đã bán được gì tại Hải Phòng?	Đã bán được gì trong năm ngoái, tại miền Trung Việt Nam	Sẽ bán được gì tại Cần Thơ, vào năm sau?
Năng lực về công nghệ	Lưu trữ máy tính	Cơ sở dữ liệu quan hệ	Cơ sở dữ liệu nhiều chiều và kho dữ liệu	Thuật toán nâng cao, cho phép xử lý cơ sở dữ liệu với đa bộ xử lý
Đặc tính	Phân phối tóm tắt về dữ liệu quá khứ	Phân phối dữ liệu động, theo mức bản ghi dữ liệu	Phân phối dữ liệu quá khứ theo nhiều mức độ	Chủ động phân phối thông tin



Hình 1.9. Vai trò của khai phá dữ liệu

1.2.8. Thách thức chính của khai phá dữ liệu

1.2.8.1. Các thách thức

Một số thách thức chính trong dự án khai phá dữ liệu là:

- *Sử dụng* dữ liệu trong cơ sở dữ liệu giao tác để khai phá dữ liệu;
- *Giảm chiều* dữ liệu;
- *Chuyển hóa* dữ liệu;
- *Làm sạch* dữ liệu;
- *Xử lý dữ liệu thưa*;
- *Dữ liệu hiếm*, tức các khuôn mẫu ít thấy, liên quan đến việc làm cân bằng bộ dữ liệu.

Mỗi thách thức này cần được trao đổi chi tiết.

1.2.8.2. Một số thí dụ về ứng dụng

Công nghệ khai phá dữ liệu được ứng dụng bất kì nơi cần ra quyết định, dựa trên các sự kiện. Có một số loại ứng dụng;

- *Dự đoán bán hàng*. Đây là ứng dụng sớm nhất của khai phá dữ liệu;
- *Quản lí giá hàng*. Nó tiếp tục ứng dụng về dự đoán bán hàng, để quản lí các giá đặt hàng hóa trong kho, siêu thị...;
- *Khám phá khoa học*. Người ta dùng khai phá dữ liệu để định danh các đối tượng vũ trụ;
- *Trò chơi*. Phương pháp dự đoán giúp người chơi đạt kì lục về điểm;
- *Thể thao*. Sử dụng công cụ khai phá dữ liệu để lựa chọn tình huống trong thể thao, về tỉ số bàn thắng, cầu thủ...;
- *Quản trị quan hệ khách hàng*. Việc này xác định xu hướng kinh doanh;

- *Thu thập khách hàng.* Người ta dùng khai phá dữ liệu để định danh khả năng tương lai trong kinh doanh.

1.2.8.3. Vai trò của tri thức

Một phân tích khai phá dữ liệu xây dựng mô hình với bộ dữ liệu và phát hiện tính dự đoán thấp trong kết quả. Phân tích khác có thể bắt đầu với cùng bộ dữ liệu, nhưng tạo được mô hình kết quả có tính dự đoán cao. Điều khác biệt là do khâu chuẩn bị dữ liệu, chứ không phải khâu lựa chọn thuật toán.

Định nghĩa: Tri thức là hiểu biết về lĩnh vực.

Một vài thuật toán chỉ thích hợp với một số loại bộ dữ liệu. Người phân tích hiểu rõ về lĩnh vực kinh doanh, nên lựa chọn được thuật toán phù hợp, để áp lên bộ dữ liệu thu thập. Điều này cho thấy cần có tri thức trong quá trình khai phá dữ liệu. Có nhiều dạng tri thức, từ nhiều nguồn khác nhau, đáp ứng việc trợ giúp quá trình khai phá dữ liệu.

1.2.9. Các khía cạnh chính trong khai phá dữ liệu

Sau đây là một số khía cạnh chính trong khai phá dữ liệu:

1. *Khai phá các loại thông tin khác nhau trong cơ sở dữ liệu.* Cần tích hợp dữ liệu từ nhiều nguồn khác nhau, cả kho dữ liệu, bảng tính điện tử... Việc tích hợp là phức tạp và tốn thời gian;
2. *Khai phá tương tác của tri thức tại mọi cấp độ.* Dữ liệu mức tính toán cần kết hợp với dữ liệu mức cá nhân, theo thang độ thời gian đa dạng, tức theo ngày, hay tuần... Khía cạnh này đòi hỏi chuyển hóa cẩn thận các kiểu dữ liệu đầu vào;
3. *Kết hợp thông tin nên.* Một vài biến lấy từ ngoài cơ sở dữ liệu. Chúng có thể là dữ liệu dân số, dữ liệu lịch sử hay từ bên thứ ba. Cần đảm bảo chính xác và phù hợp khi tích hợp dữ liệu này với dữ liệu trong cơ sở dữ liệu;

4. *Ngôn ngữ hỏi trong khai phá dữ liệu và khai phá tùy biến.* Người khai phá cần tương tác với hệ quản trị cơ sở dữ liệu để truy cập dữ liệu. Người ta dùng câu SQL, nhưng cũng cần biết kỹ năng về đại số quan hệ hay lập trình;
5. *Thể hiện và hiển thị kết quả khai phá.* Vấn đề là giao diện cho phép hiển thị kết quả kỹ thuật cho người dùng không chuyên công nghệ thông tin, với khả năng thân thiện và có tương tác. Đồ họa có ý nghĩa hơn con số trong giải tích;
6. *Quản lý nhiễu và dữ liệu không đầy đủ.* Dữ liệu không thể hoàn toàn đáp ứng mô hình, nên hiện trạng: thiếu, sai, lệch, không chắc chắn, mờ, rỗng... về dữ liệu cũng xảy ra trong khai phá dữ liệu;
7. *Đánh giá khuôn mẫu.* Nhiều khuôn mẫu có trong bộ dữ liệu. Thách thức đối với quá trình khai phá là phân biệt các khuôn mẫu này, để lựa chọn khuôn mẫu có ý nghĩa, đáng quan tâm. Như vậy sẽ có độ đo về mức độ phù hợp;
8. *Tính hiệu quả và khả năng mở rộng của thuật toán khai phá.* Hiệu quả của thuật toán khai phá dữ liệu được đo theo năng lực dự đoán và thời gian sinh ra mô hình. Khía cạnh mở rộng đặt ra khi đưa thuật toán hay mô hình xây dựng được trên bộ dữ liệu nhỏ áp dụng trên bộ dữ liệu lớn hơn. Các thuật toán và mô hình đều mở rộng được một cách tuyến tính; tức thời gian xử lý tăng theo hình học, theo vì tăng theo hàm mũ của kích thước bộ dữ liệu;
9. *Các thuật toán gia tăng, phân tán và song song.* Bài toán khai phá lớn được chia nhỏ rồi xử lý, trên nhiều bộ xử lý, trên máy tính song song. Cũng có bài toán thực hiện với thuật toán gia tăng, với tính toán lưới, trên dòng dữ liệu liên tục, thay vì trên các ngăn dữ liệu;
10. *Quản lý kiểu dữ liệu quan hệ và phức tạp.* Nhiều dữ liệu đầu vào lấy từ cơ sở dữ liệu quan hệ. Cũng có thể dữ liệu lấy từ cơ sở dữ liệu đa chiều phức tạp. Quá trình khai phá dữ liệu cần linh hoạt để thực hiện với cả hai loại dữ liệu;

11. *Khai phá thông tin từ hệ thống thông tin tổng quát, không đồng nhất.* Hệ thống không đồng nhất gồm nhiều kiểu dữ liệu, nhiều mô hình, trên hạ tầng kỹ thuật đa dạng. Quá trình khai phá cần thực hiện trong điều kiện ấy, nên giao diện người dùng đồ họa GUI¹ có vai trò quan trọng.

1.2.10. Yêu cầu chung để thành công trong dự án khai phá dữ liệu

1.2.10.1. Một số yêu cầu về dự án

Để đi đến đích của quá trình khai phá, cần đảm bảo một số yêu cầu. Các yêu cầu này ở mức tổng quát; tùy tình huống cụ thể, mà người ta có yêu cầu riêng, đối với thuật toán và kinh nghiệm người khai phá phù hợp...

1. *Mong kết quả có ý nghĩa.* Thông thường, kết quả của khai phá dữ liệu (i) đáp ứng đích, do người ta đã chuẩn bị dữ liệu và thuật toán, mô hình tốt; (ii) có thể cải thiện kết quả, tức nếu tăng cường khai phá, kết quả sẽ tốt hơn; tuy nhiên dừng ở mức trung gian bởi lẽ đó là kết quả chấp nhận được và không cần tốn phí thêm;
2. *Đội công tác có kỹ năng đối với hoạt động yêu cầu.* Đối với dự án, có nhiều chức vụ quản lý khác nhau. Việc thành lập đội công tác, vai trò của người lãnh đạo có ý nghĩa đối với thành công của hoạt động dự án. Quan trọng hơn, nhân viên phân tích cần hợp tác với người kinh doanh, để chuyên gia phân tích có thể làm việc với tri thức lĩnh vực;
3. *Cảnh giác với dữ liệu.* Khi thu lượng dữ liệu thủ công hay tự động, cần xử lý dữ liệu sơ bộ trước khi tích lũy trong cơ sở dữ liệu;
4. *Thời gian.* Quá trình học được lặp lại, nên cần có thời gian. Mặt khác công việc dự án thường bị trễ, tác động xấu lên hoạt động khác.

¹ GUI Graphical User Interface: giao diện người dùng đồ họa.

Các mô hình khai phá dữ liệu cần chạy đua với thực tế. Mô hình bổ sung khách hàng dùng trong đợt tiếp thị đầu tiên là chưa tốt ngay, nhưng sẽ thu được kết quả tốt khi lặp lại liên tiếp.

1.2.10.2. Lí do cần đến khai phá dữ liệu

Đã một thời gian dài người ta quen với phân tích thống kê. Khi khai phá dữ liệu xen vào là lúc người ta nhận thấy tính cần thiết, có nhu cầu. *Tính cấp thiết là cha của phát minh.* Trong trên 50 năm qua, công nghiệp, kinh doanh, xã hội đã tích tụ nhiều dữ liệu. Người ta còn đánh giá rằng trên 90% tri thức là thu được trong khoảng 50 năm đó.

Đối mặt với khối dữ liệu lớn, một dạng học mới hình thành. Tiếp cận mới để ra quyết định dựa trên dữ liệu đầu vào đã dựng nên môi trường mới bên khối dữ liệu đó. Chuyên ngành trí tuệ nhân tạo đề xuất sử dụng mô hình não người thay cho mô hình tham số của Fisher. Ngay từ lúc đầu, mạng nơron đã phát triển quá trình tương tự con người, và cây quyết định, với hệ thống phân cấp của khẳng định có/không, đã phát triển như một tiếp cận hệ thống, để khám phá sự thật trong thế giới.

Các tiếp cận khai phá dữ liệu cũng được ứng dụng cho bộ dữ liệu tương đối nhỏ, với độ chính xác dự đoán như các kĩ thuật thống kê, hoặc tốt hơn.

1.2.10.3. Một vài cảnh báo với giải pháp của khai phá dữ liệu

Bên cạnh những thành công từ khai phá dữ liệu, người ta cũng cần nhận thức được mặt trái của nó. Sau đây là một vài cảnh báo về sử dụng các công cụ khai phá dữ liệu để khám phá khuôn mẫu.

1. *Chất lượng dữ liệu.* Chất lượng dữ liệu tồi có thể không lộ ra tường minh trong các phương pháp khai phá dữ liệu, và sẽ cho ra mô hình tồi. Cũng có thể dữ liệu tồi đó được dùng để xây dựng mô hình với mức dự đoán tương đối cao, nhưng đó là ảo tưởng;

2. *Cơ hội*. Nhiều cơ hội có thể chuyển những thứ không thể thành sự kiện có thể. D. Hand, năm 2005, đã đề cập vấn đề này như *vấn đề đa dạng*;
3. *Can thiệp*. Có khi thu được kết quả không lường trước, khiến người ta cần thay đổi, chỉnh lí để làm mất hiệu lực của kết quả không mong muốn. Hệ thống đã học và có biện pháp ngăn ngừa xâm nhập trái phép; tuy nhiên tin tặc có thể can thiệp để tiếp tục phá hoại;
4. *Tách biệt*. Thường thì khóa tách thông tin quan tâm khỏi đồng thông tin trong bộ dữ liệu. Nhiều khuôn mẫu đã tồn tại trong bộ dữ liệu, nhưng chỉ một ít khuôn mẫu là đáng quan tâm đối với người dùng trong vấn đề đang xét. Do vậy việc xác định biến đích là nhân tố quan trọng, để thuật toán biết khuôn mẫu cần thiết;
5. *Rõ ràng*. Một vài khuôn mẫu trong bộ dữ liệu có thể không bao giờ dùng đến, do nó hoàn toàn rõ ràng, hiển nhiên, không cần thiết phân tích thống kê;
6. *Không ổn định*. Tính mất ổn định xảy ra khi quá trình tạo nên thay đổi trên bộ dữ liệu, theo ý riêng của mình.

1.3. Một số chủ đề liên quan đến khai phá dữ liệu

Trong cuốn *Cẩm nang về khai phá dữ liệu* của Nong Ye (2003), các chủ đề liên quan đến *khai phá dữ liệu* gồm:

Thứ nhất là *phương pháp luận về khai phá dữ liệu*:

1. Cây quyết định;
2. Các luật kết hợp;
3. Mô hình mạng nơron nhân tạo;
4. Phân tích thống kê về dữ liệu bình thường và bất thường;
5. Phân tích dữ liệu Bayes;
6. Quá trình Markov ẩn và khai phá mẫu tuần tự;

7. Các chiến lược và phương pháp dự đoán;
8. Các thành phần chính và phân tích nhân tố;
9. Các phương pháp đo tâm lí của mô hình hóa biến ngầm định;
10. Phân cụm mờ rộng;
11. Đo tương tự chuỗi thời gian và chỉ mục hóa;
12. Phân tích chuỗi thời gian phi tuyến;
13. Khai phá dữ liệu phân tán.

Thứ hai là *việc quản trị khai phá dữ liệu*:

1. Thu thập, chuẩn bị, chất lượng và hiển thị dữ liệu;
2. Lưu trữ dữ liệu và quản lí;
3. Trích đặc trưng, lựa chọn và kiến thiết;
4. Phân tích hiệu năng và đánh giá;
5. An toàn và tính riêng tư;
6. Các chuẩn mới nổi và giao diện.

Thứ ba là *các ứng dụng của khai phá dữ liệu*:

1. Khai phá dữ liệu về hiệu năng con người;
2. Khai phá dữ liệu văn bản;
3. Khai phá dữ liệu địa lí;
4. Khai phá dữ liệu khoa học và công nghệ;
5. Khai phá dữ liệu trong tin sinh;
6. Khai phá dữ liệu quản trị quan hệ khách hàng;
7. Khai phá dữ liệu an toàn mạng và máy tính;
8. Khai phá dữ liệu hình ảnh;
9. Khai phá dữ liệu chất lượng sản xuất.

Thông thường, người thực hiện việc khai phá dữ liệu phải trả lời về kết quả phân tích thống kê và khai phá dữ liệu. Người ta cần biết khai phá dữ liệu ra sao theo quan điểm thủ tục.

Phân tích thống kê và khai phá dữ liệu là hai phương pháp để mô phỏng các phép toán vô thức, thường xảy ra trong não người, để đảm bảo ra quyết định và hành động theo cách hợp lí. (i) Phân tích thống kê là suy luận trực tiếp, dựa trên các quy tắc. Người ta suy nghĩ và quyết định trên cơ sở của quy tắc. Sử dụng các khái niệm tương tự về trung bình và độ lệch chuẩn, các phân tích thống kê tiến hành chứng minh theo cách hợp lí; (ii) Khai phá dữ liệu học từng trường hợp, không sử dụng trung bình và độ lệch chuẩn. Các thuật toán khai phá dữ liệu xây dựng các mẫu, làm rõ mẫu trong các trường hợp cần xử lí. Hai cách khác nhau xa lại cùng nhắm một đích: quyết định.

1.3.1. Mẫu hành động

Mô hình hóa thống kê là việc giống như con người cần hoạt động trong khuôn khổ mẫu; vừa khó vượt qua khuôn khổ, nhưng muốn tháo bỏ các ràng buộc.

Định nghĩa: Mẫu hành động là dạng hoạt động được hạn chế với phạm vi thích hợp.

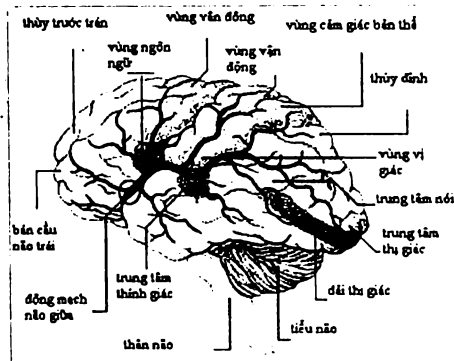
Việc huấn luyện mô hình hành động thực hiện theo một loạt các bước (i) thử dần dần; (ii) thử di chuyển ra khỏi ràng buộc; (iii) quyết định cách tốt nhất để thực hiện các hoạt động; (iv) thử các cách tốt nhất đối với các hoạt động.

Việc huấn luyện như vậy phù hợp với hệ thống cảm biến và các tế bào thần kinh. Tế bào của não được huấn luyện bằng cách thử sức bền chịu đựng và vượt qua ngưỡng.

1.3.2. Trực giác của người

Não trái của con người điều khiển các hoạt động logic, định lượng; não phải điều khiển các hoạt động nhạy cảm với thẩm mỹ và hiển thị. Có dây thần kinh nối hai bán cầu não; số dây này của nữ giới cao hơn nam giới 20%. Điều này chứng tỏ phụ nữ thuận về hoạt động trực giác hơn đàn ông.

Khả năng trực giác cho phép xây dựng mô hình hiện thực từ (i) thành phần định lượng; và (ii) nhạy cảm hiện thị.



Hình 1.10. Não người

Con người có khả năng (i) phân tích; (ii) tổng hợp, với năng lực (1) cụ thể hóa; (2) tổng quát hóa. Nhiều hoạt động của người liên quan đến phát hiện mẫu trong dữ liệu nhập vào, qua hệ thống các cảm biến. Cần kết luận từ mẫu này, dù mẫu xuất hiện trong lĩnh vực nhỏ. Cả phân tích thống kê và khai phá dữ liệu đều thao tác trên các mẫu. (i) Phân tích thống kê sử dụng các mẫu xác định trước, tức theo mô hình tham số, và so sánh độ đo nào đó của quan sát theo độ đo chuẩn của mô hình; (ii) Khai phá dữ liệu không bắt đầu với mô hình, mà *xây dựng* mô hình với dữ liệu. Phân tích thống kê dùng mô hình để đặc trưng một mẫu trong dữ liệu; khai phá dữ liệu dùng mẫu này trong dữ liệu để xây dựng mô hình. Tiếp cận này theo *lập luận suy diễn*.

Liên quan đến *phân tích thống kê và khai phá dữ liệu*, người ta đề cập:

Thứ nhất là *lịch sử* các pha phân tích dữ liệu, lý thuyết cơ bản và quá trình khai phá. Chi tiết gồm:

1. Nền tảng của ứng dụng khai phá dữ liệu;
2. Khía cạnh lý thuyết của khai phá dữ liệu;

3. Quá trình khai phá dữ liệu;
4. Hiểu và chuẩn bị dữ liệu;
5. Lựa chọn đặc trưng;
6. Các công cụ cho phép khai phá dữ liệu.

Thứ hai là các *thuật toán* khai phá dữ liệu và khai phá văn bản:

1. Các thuật toán cơ bản cho phép khai phá dữ liệu;
2. Khai phá văn bản và xử lý ngôn ngữ tự nhiên;
3. Công cụ phần mềm thông dụng nhất.

Thứ ba là những *ứng dụng chính* của khai phá dữ liệu:

1. Phân lớp;
2. Dự đoán số học;
3. Đánh giá mô hình;
4. Tin học y tế;
5. Tin sinh;
6. Mô hình hóa việc trả lời khách hàng;
7. Phát hiện lỗi;

Thứ tư là một số *vấn đề liên quan* đến việc khai phá dữ liệu. Đó là việc đo độ phức tạp, xác định mô hình sử dụng tốt, các lỗi thường mắc và xu thế phân tích:

1. Độ phức tạp mô hình;
2. Mô hình đúng đối với mục đích đúng;
3. Các lỗi về khai phá dữ liệu;
4. Tương lai về khai phá dữ liệu và khai phá văn bản.

Trong tài liệu về khai phá dữ liệu năm 2014, quan niệm về chuyên ngành này không khác nhiều. Về các khái niệm cơ bản của *phân tích dữ liệu*, người ta đề cập (i) khai phá dữ liệu và phân tích; (ii) các thuộc tính số; (iii) các thuộc tính phạm trù; (iv) dữ liệu đồ

thị; (v) các phương pháp nhân; (vi) dữ liệu nhiều chiều; và (vi) việc giảm chiều. Liên quan đến *khai phá các mẫu thường xuyên*, có các chủ đề (i) khai phá các tập mục; (ii) tóm tắt các tập mục; (iii) khai phá chuỗi; (iv) khai phá mẫu đồ thị; (v) mẫu và đánh giá luật. Về việc *phân cụm*, người ta đề cập (i) phân cụm dựa trên đại diện; (ii) phân cụm phân cấp; (iii) phân cụm dựa trên mật độ; (iv) phân cụm đồ thị và phổ; (v) kiểm chứng việc phân cụm. Về việc *phân lớp*, có những chủ đề (i) phân lớp xác suất; (ii) phân lớp dùng cây quyết định; (iii) phân tích dị biệt tuyến tính; (iv) máy vectơ tựa.

1.3.3. Bài toán khai phá dữ liệu

Bài toán khai phá dữ liệu trong các chương liên quan đến (i) khai phá luật kết hợp; (ii) phân lớp; (iii) phân cụm. Để thực hiện việc này, cần có (i) phân tích thống kê; (ii) các kỹ thuật trợ giúp, chẳng hạn OLAP, OLTP...; (iii) các mô hình được xây dựng dựa trên dữ liệu.

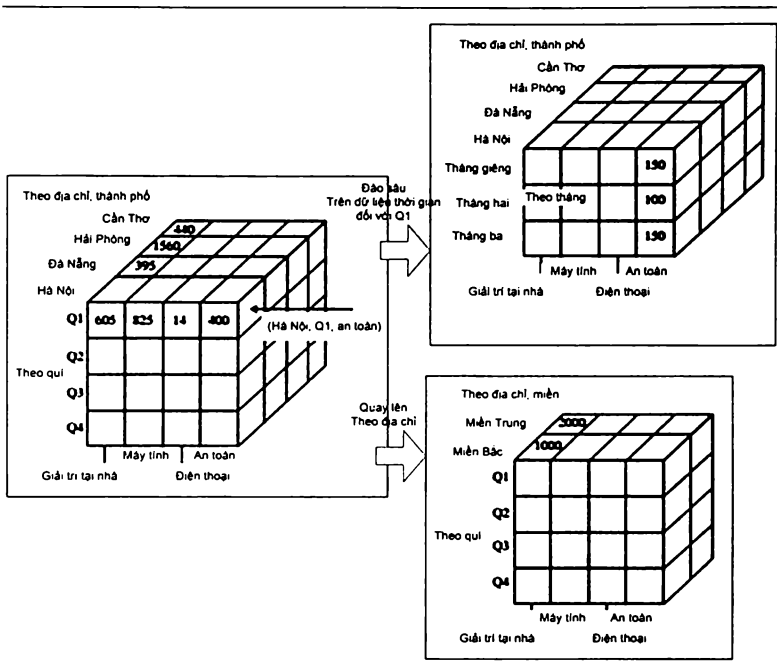
Thí dụ được sử dụng theo tài liệu tham chiếu (i) bài toán phân tích tại cửa hàng đồ điện; (ii) bài toán phân tích dữ liệu về hoa diên vĩ.

1.3.3.1. Mẫu có ích

Người ta thường quan niệm sai lầm về khai phá dữ liệu là mong muốn rằng hệ thống khai phá dữ liệu có thể tự động phát hiện ra tất cả các tri thức có giá trị có trong một cơ sở dữ liệu lớn, mà không cần sự can thiệp hoặc sự chỉ dẫn của con người.

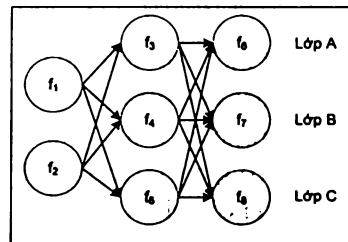
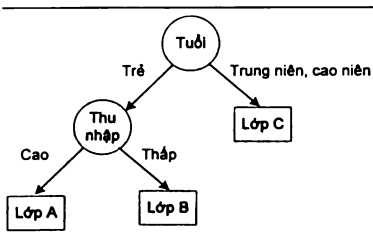
Mặc dù hấp dẫn đầu tiên là hệ thống khai phá dữ liệu tự động, nhưng trong thực tế, những hệ thống như vậy sẽ khám phá ra một tập *mẫu*¹ lớn. Toàn bộ tập mẫu được sinh ra có lẽ dễ dàng vượt quá kích thước của cơ sở dữ liệu cần thiết.

¹ Pattern: mẫu



Hình 1.11. Khối dữ liệu đa chiều, thường dùng với kho dữ liệu (a) tổng hợp dữ liệu về Hàng điện tử; (b) đào sâu... thực hiện trên các khối dữ liệu

Tuổi (X, trẻ) ^ thu nhập (X, cao) → Lớp (X, A)
 Tuổi (X, trẻ) ^ thu nhập (X, thấp) → Lớp (X, B)
 Tuổi (X, trung niên) → Lớp (X, C)
 Tuổi (X, cao niên) → Lớp (X, D)



Hình 1.12. Mô hình phân lớp thể hiện qua nhiều dạng (a) luật IFTHEN; (b) cây quyết định; (c) mạng nơron.



Hình 1.13. Sinh ra nhiều dữ liệu do khai phá dữ liệu

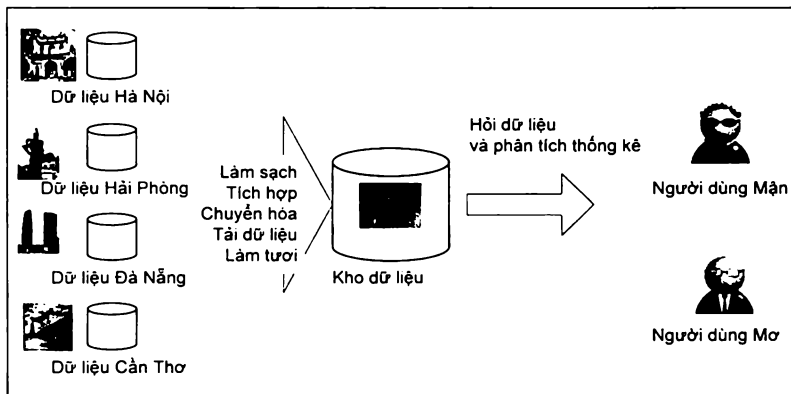
1.3.3.2. Sinh mẫu

Thuận tiện cho hệ thống là: người ta cho phép hệ thống khai phá dữ liệu thực hiện thoải mái việc khai phá các mẫu, mà không cần cung cấp cho nó bất kỳ sự chi dẫn nào. Các chi dẫn này thường là (i) nội dung cơ sở dữ liệu mà người dùng muốn thăm dò; (ii) các kiểu của mẫu mà người dùng mong muốn tìm ra. Hầu hết các mẫu được khám phá sẽ là không thích hợp để phân tích chức năng của người dùng. Hơn nữa nhiều mẫu được phát hiện mặc dù có liên quan tới chức năng phân tích nhưng không ích lợi, do (i) khó hiểu; (ii) ít giá trị, không mới lạ; (iii) thiếu tiện ích, gây nên thiếu hấp dẫn. Vì vậy khai phá dữ liệu ra các mẫu như vậy là không thuận tiện cho người khai phá dữ liệu.

1.3.3.3. Quá trình khai phá dữ liệu

Một kịch bản thực tế là: (i) người dùng có thể giao tiếp với hệ thống khai phá dữ liệu; (ii) người dùng sử dụng một tập khai phá dữ liệu nguyên thủy, được thiết kế để tăng hiệu quả và khai phá tri thức có giá trị. Những dữ liệu nguyên thủy bao gồm những (i) đặc tả của các phần trong cơ sở dữ liệu, hoặc bộ dữ liệu mà người dùng quan tâm; (ii) các thuộc tính cơ sở dữ liệu, hoặc kích thước của kho dữ liệu quan tâm. Các loại kiến thức được khai thác, kiến thức nền hữu ích trong quy trình khai phá, các phương pháp hiệu quả để đánh giá mẫu và tri thức khai phá nên được thể hiện trực

quan. Những phép toán nguyên thủy cho phép người dùng giao tiếp trực tiếp với hệ thống khai phá dữ liệu trong suốt quá trình khai phá để nghiên cứu, phát hiện ra từ các góc hoặc chiều sâu và quy trình khai phá trực tiếp.



Hình 1.14. Khung điển hình đối với kho dữ liệu, bài toán Hàng điện tử

1.3.3.4. Ngôn ngữ khai phá

Ngôn ngữ hỏi khai phá dữ liệu có thể được thiết kế để kết hợp các phép toán nguyên thủy cho phép người dùng tương tác linh hoạt với các hệ thống khai phá dữ liệu. Ngôn ngữ hỏi khai phá dữ liệu cung cấp một nền tảng giao diện đồ họa người dùng thân thiện. Trong chương này, có thể tìm hiểu về các *phép toán nguyên thủy khai phá dữ liệu* một cách chi tiết cũng như nghiên cứu cách *thiết kế của ngôn ngữ hỏi khai phá dữ liệu* dựa trên các nguyên tắc này.

<i>customer</i>	(<i>cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...</i>)
<i>item</i>	(<i>item_ID, brand, category, type, price, place_made, supplier, cost, ...</i>)
<i>employee</i>	(<i>empl_ID, name, category, group, salary, commission, ...</i>)
<i>branch</i>	(<i>branch_ID, name, address, ...</i>)
<i>purchases</i>	(<i>trans_ID, cust_ID, empl_ID, date, time, method_paid, amount</i>)
<i>items_sold</i>	(<i>trans_ID, item_ID, qty</i>)
<i>works_at</i>	(<i>empl_ID, branch_ID</i>)

Hình 1.15. Lược đồ cơ sở dữ liệu thí dụ về hàng điện tử, với chú thích tiếng Anh

Cơ sở dữ liệu thí dụ là *Hàng điện tử*. Cơ sở dữ liệu có các quan hệ sau:

- Khách hàng (#khách hàng, tên, địa chỉ, tuổi, nghề, thu nhập, thông tin tài chính, lĩnh vực)
- Hàng (#hàng, chi nhánh, lĩnh vực, loại, giá, nơi sản xuất, nhà cung cấp)
- Nhân viên (#nhân viên, tên, lĩnh vực, nhóm, lương, nhiệm vụ)
- Chi nhánh (#chi nhánh, tên, địa chỉ)
- Mua (#giao dịch, #khách hàng, #nhân viên, ngày, giờ, phương thức thanh toán, tổng)
- Hàng bán được (#giao dịch, #hàng, số lượng)
- Làm việc (#nhân viên, #chi nhánh)

1.4. Yêu cầu đối với dữ liệu trong bài toán thí dụ

Phần này tiếp tục đề cập yêu cầu đối với bài toán thí dụ trong một số chương. Đây là các yêu cầu liên quan đến dữ liệu của bài toán.

Định nghĩa: Nhiệm vụ khai phá dữ liệu đối với mỗi người dùng là nhiệm vụ trong mỗi người dùng, có các chức năng (i) phân tích thống kê dữ liệu; (ii) thể hiện dữ liệu.

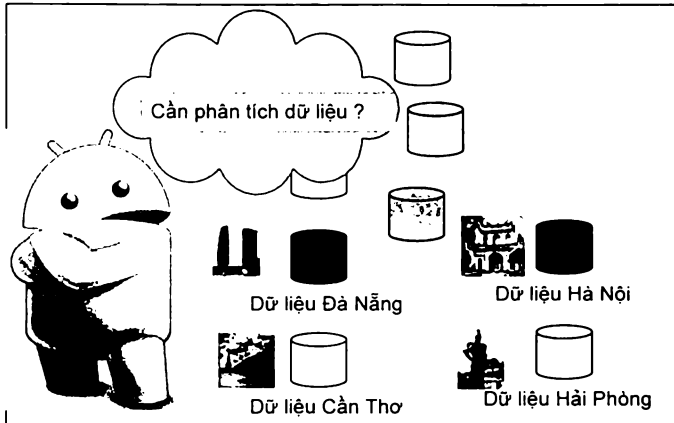
Ngôn ngữ hỏi khai phá dữ liệu được định nghĩa theo các thuật ngữ (i) dữ liệu cho quá trình khai phá dữ liệu; (ii) loại truy vấn; (iii) tri thức nền; (iv) công cụ đánh giá, đo kết quả; (v) cách phát hiện mẫu.

Định nghĩa: Ngôn ngữ khai phá dữ liệu là ngôn ngữ xác định đầu vào, đầu ra của hệ thống khai phá, qua các đặc tả của ngôn ngữ.

1.4.1. Các nhiệm vụ cần thực hiện

Bài toán thí dụ trong phần này là quản lý bán hàng. Người ta muốn nghiên cứu về xu hướng mua hàng của các khách hàng. Thay vì khai thác trên toàn bộ cơ sở dữ liệu, người ta có thể xem xét chỉ di

liệu liên quan đến việc mua hàng của khách hàng. Chúng thể hiện cùng với thông tin khách hàng liên quan.



Hình 1.16. Nhiều thông tin, yêu cầu phân tích thống kê

Có thể chỉ ra các thuộc tính quan tâm để được xem xét trong quá trình khai phá. Những thuộc tính này gọi là thuộc tính liên quan. Chẳng hạn chỉ quan tâm tới việc nghiên cứu giữa các mối quan hệ, các mặt hàng được mua, lương và tuổi của khách hàng hàng năm, các thuộc tính *tên* của quan hệ *hàng*, *thu nhập* và *tuổi* của quan hệ *khách hàng* có thể được đặc tả như là các thuộc tính thích hợp cho quá trình khai phá. Phần cơ sở dữ liệu được khai phá tạo nên khung nhìn dữ liệu. Khung nhìn gồm một nhóm của một hoặc nhiều tập các thuộc tính hay các chiều.

1.4.2. Các loại tri thức được khám phá

1.4.2.1. Phân loại yêu cầu về tri thức

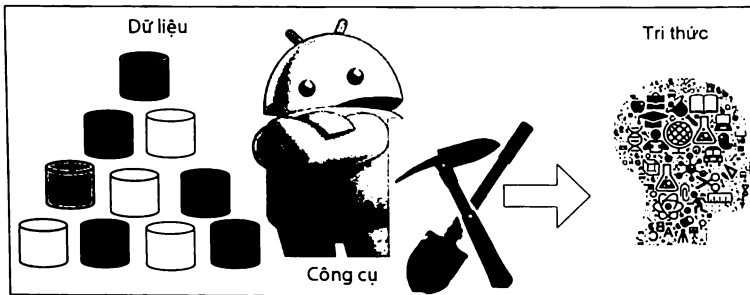
Việc xác định các loại tri thức được khám phá là rất quan trọng. Vì điều này quyết định chức năng khai phá dữ liệu được thực hiện. Có nhiều phân loại về tri thức; ở đây đề cập các tri thức (i) mô tả khái niệm, theo đặc tính và sự mô tả; (ii) sự kết hợp; (iii) sự phân lớp, phân nhóm; (iv) dự đoán và phân tích thống kê, phân tích phát triển.

Định nghĩa: Khám phá tri thức, hay trích rút tri thức¹, là việc tạo kiến thức từ các cơ sở dữ liệu có cấu trúc, như cơ sở dữ liệu quan hệ, XML, từ tập không có cấu trúc, như văn bản, tài liệu, hình ảnh.

Ngoài ra việc xác định các kiểu tri thức được khám phá có nhiệm vụ khai phá dữ liệu được đưa ra, người dùng có thể đặc và đưa ra các *dạng mẫu*² mà các mẫu được khai phá phải phù hợp Các khuôn mẫu hoặc các siêu mẫu, hay được gọi là siêu luật, hay siêu câu hỏi, có thể dùng để hướng dẫn quá trình khám phá.

1.4.2.2. Kỹ thuật khám phá

Tri thức được khám phá cho thấy kết quả của các chức năng của khai phá dữ liệu; đó là các đặc tính, nhận thức, sự kết hợp, phân lớp, sự phân nhóm hoặc phân tích quá trình tiến hóa.



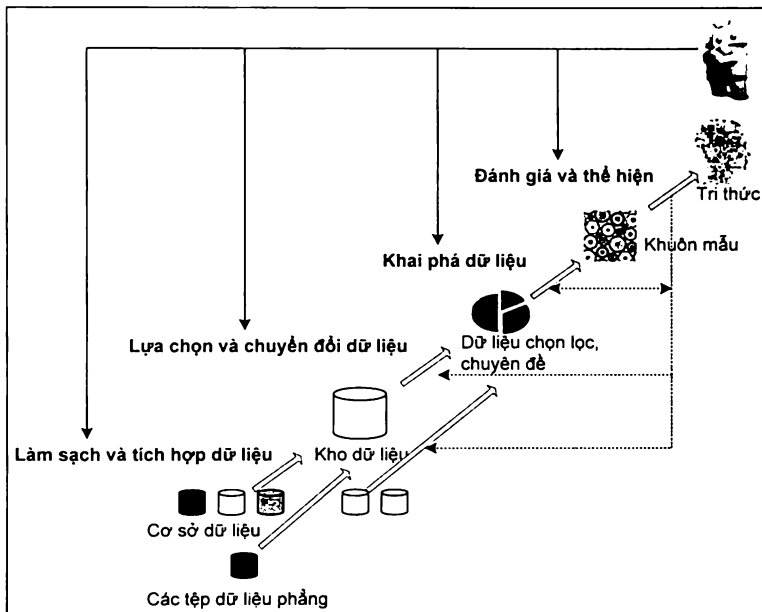
Hình 1.17. Tri thức rút ra từ dữ liệu

- **Tri thức nền tảng.** Người dùng có thể xác định kiến thức nền hoặc kiến thức thuộc lĩnh vực khai thác. Các kiến thức này có ích trong việc hướng dẫn quá trình khai phá tri thức và đánh giá các mẫu được phát hiện. Có nhiều kiểu kiến thức nền. Ở đây đề cập kiểu kiến thức nền thông dụng nhất, là *phân cấp khái niệm*. Phân cấp khái niệm rất thích hợp cho phép dữ liệu được khai phá ở nhiều mức độ.

¹ Knowledge discovery, knowledge extract: khám phá, trích rút tri thức.

² Template: dạng mẫu (để phân biệt với mẫu, sample, và khuôn mẫu, pattern).

độ trừu tượng; chẳng hạn niềm tin của người sử dụng về các mối quan hệ trong dữ liệu. Điều này có thể được dùng để đánh giá các mẫu được khai phá theo tình huống đột xuất, bất ngờ khi các mẫu không mong muốn được cho là hấp dẫn;



Hình 1.18. Khai phá dữ liệu như một bước trong khám phá tri thức

- *Các kỹ thuật, chức năng.* Các chức năng này được sử dụng để tách các mẫu không mong muốn từ tri thức. Chúng được sử dụng để hướng dẫn quá trình khai phá hoặc sau khi khai phá, nhằm đánh giá các mẫu khai phá. Các kiểu tri thức khác nhau dùng cho phương pháp khác nhau. Chẳng hạn các phương pháp cho các luật kết hợp, với các trợ giúp, tức phần trăm của dữ liệu thích hợp mà trong luật mẫu xuất hiện, và độ tin cậy, tức độ mạnh của hàm ý trong luật. Các luật có giá trị trợ giúp và giá trị tin cậy dưới ngưỡng của người dùng đặc tả sẽ được xem như không hấp dẫn;

- *Trình diễn và trực quan hóa các mẫu được khai phá.* Điều này hàm ý về dạng hiển thị của các mẫu được khai phá. Người sử dụng có thể chọn các dạng khác nhau để trình diễn kết quả như là các bảng luật, biểu đồ, đồ thị, cây quyết định và hình khối.

1.4.3. Dữ liệu liên quan tới nhiệm vụ

Dữ liệu liên quan đến nhiệm vụ khai phá dữ liệu là phần quan trọng trong toàn cơ sở dữ liệu, cần được nghiên cứu tỉ mỉ. Điều quan tâm đầu tiên là đặc tả về dữ liệu cần khai phá. Thông thường, người sử dụng chỉ quan tâm một tập con của cơ sở dữ liệu. Việc khai phá trên toàn cơ sở dữ liệu là không thực tế. Lí do là số lượng mẫu sinh ra có thể là cấp số nhân so với kích thước của cơ sở dữ liệu. Hơn nữa các mẫu được phát hiện có thể là không thích hợp với quan tâm của người dùng.

1.4.3.1. Dữ liệu thích hợp với nhiệm vụ

Dữ liệu thích hợp với nhiệm vụ khai phá dữ liệu theo các góc nhìn:

- Cơ sở dữ liệu hoặc tên kho dữ liệu;
- Các bảng cơ sở dữ liệu hoặc các khối trong kho dữ liệu;
- Các điều kiện lựa chọn dữ liệu;
- Các chiều hoặc các thuộc tính thích hợp;
- Các điều kiện nhóm các dữ liệu.

1.4.3.2. Kiểu tri thức được khai phá

Liên quan đến tri thức cần khám phá, người ta quan tâm đến:

- Đặc tính, tính chất của dữ liệu;
- Quá trình suy diễn, nhận thức;
- Các mối quan hệ, tương quan;
- Yêu cầu về phân lớp, phân nhóm, dự đoán...

1.4.3.3. Kiến thức nền tảng

Một số kiến thức cần cho các bài toán thí dụ:

- Phân cấp khái niệm;
- Tin cậy của người sử dụng về các mối quan hệ trong dữ liệu.

1.4.3.4. Độ đo liên quan đến hấp dẫn của mẫu

Đối với các mẫu, có các độ đo sau:

- Tính đơn giản;
- Tính chắc chắn, tự tin;
- Tính tiện dụng, hay tiện ích;
- Tính mới.

1.4.3.5. Trục quan hóa các mẫu được khai phá

Trong một cơ sở dữ liệu quan hệ, có thể tìm được bộ dữ liệu liên quan tới chức năng nhờ câu hỏi liên quan tới các thao tác (i) chọn, tức hạn chế; (ii) chiếu; (iii) nối, tức nhân... Việc truy cập dữ liệu có thể được xem như nhiệm vụ con trong nhiệm vụ khai phá dữ liệu. Kết quả của quá trình phát hiện dữ liệu trong dữ liệu cho phép tạo nên *quan hệ dữ liệu mở rộng*. Quan hệ dữ liệu có thể được sắp xếp hoặc nhóm lại theo điều kiện trong câu hỏi. Dữ liệu có thể được làm sạch hoặc chuyển đổi; chẳng hạn kết hợp trên các thuộc tính nào đó, ưu tiên để áp dụng cho việc phân tích khai phá dữ liệu... Quan hệ khởi tạo có thể phù hợp hoặc không với quan hệ vật lý trong cơ sở dữ liệu. Vì vậy các quan hệ ảo được gọi là *khung nhìn* trong lĩnh vực cơ sở dữ liệu; bộ dữ liệu thích hợp cho khai phá dữ liệu được gọi là *khung nhìn nhỏ*.

1.4.3.6. Thí dụ về tổ chức cơ sở dữ liệu

Nếu nhiệm vụ khai phá dữ liệu là nghiên cứu những mối quan hệ giữa các mặt hàng thường xuyên được mua tại công ty *Hàng điện tử*. Dữ liệu thích hợp có thể được đặc tả bằng việc

cung cấp các thông tin sau, như lược đồ cơ sở dữ liệu của bà toán thí dụ.

- Tên của cơ sở dữ liệu hoặc kho dữ liệu được sử dụng, thí dụ *Hàng điện tử*;
- Tên các bảng hoặc các khối dữ liệu chứa dữ liệu thích hợp chẳng hạn các bảng *hàng*, *khách hàng*, *mua*, *hàng bán được*...;
- Các điều kiện để chọn dữ liệu phù hợp, chẳng hạn tìm kiếm dữ liệu liên quan đến việc mua, ở năm hiện tại;
- Các thuộc tính hoặc các chiều liên quan, chẳng hạn *tên*, *giới tính* trong bảng *hàng*; *thu nhập* và *tuổi* trong bảng *khách hàng*;
- Ngoài ra người dùng có thể yêu cầu dữ liệu phát hiện được s nhóm theo thuộc tính cụ thể; chẳng hạn *nhóm theo thời gian*.
Câu SQL¹ cho thông tin, để tìm kiếm dữ liệu thích hợp.

Trong một kho dữ liệu, dữ liệu có thể được lưu trữ trong một cơ sở dữ liệu đa chiều, được biết như là một khối dữ liệu mà có thể được cài đặt sử dụng cấu trúc mảng đa chiều, một cấu trúc quan hệ, hoặc kết hợp cả hai. Bộ dữ liệu thích hợp có thể được đặt ra nhờ lọc dữ liệu dựa trên điều kiện, *lát cắt*, tức lấy dữ liệu chỉ một giá trị thuộc tính được đưa ra hoặc lát cắt; hoặc là *đào sâu*, tức trích cái chung của một vài lát cắt, của khối dữ liệu.



Hình 1.19. Đồ điện tử

¹ SQL: Structured Query Language: ngôn ngữ hỏi có cấu trúc.

Trong câu hỏi khai phá dữ liệu, các điều kiện để chọn dữ liệu có thể ở mức độ khái niệm cao hơn dữ liệu trong cơ sở dữ liệu hoặc kho dữ liệu. Chẳng hạn người dùng có thể dùng đặc tả lựa chọn trên các *mặt hàng* tại *Hàng điện tử*, theo khái niệm *type = giải trí tại nhà*, mặc dù các mục tin cá nhân trong cơ sở dữ liệu có thể không lưu trữ theo kiểu mà chỉ ở mức khái niệm thấp hơn; như là *TV¹, đầu đọc DVD²...* Một phân cấp khái niệm trên *hàng* được cụ thể hóa nhờ *giải trí tại nhà* là ở mức độ cao hơn, bao gồm các khái niệm ở mức thấp {*TV, đầu đọc DVD, ...*} có thể được sử dụng trong bộ dữ liệu thích hợp cho nhiệm vụ.

Một tập các thuộc tính thích hợp đưa ra sẽ liên quan tới các thuộc tính khác mà không được đề cập một cách rõ ràng, nhưng nên sử dụng chúng, do chúng được đưa ra theo phân cấp khái niệm hoặc chiều liên quan trong tập các thuộc tính thích hợp. Chẳng hạn tập các thuộc tính liên quan trong câu hỏi bao gồm *thành phố* Tuy nhiên thuộc tính này là một phần của các phân cấp khái niệm khác như là phân cấp khái niệm *phố < thành phố < quốc gia* đối với chiều dữ liệu *vị trí*. Trong trường hợp này các thuộc tính *phố, quốc gia* cũng nên xuất hiện trong tập các thuộc tính liên quan, do chúng thể hiện mức độ trừu tượng cao hơn và thấp hơn của *thành phố*. Điều này sẽ tạo điều kiện cho việc khai phá tri thức ở các mức độ trừu tượng nhờ *đi sâu và quay lại*.

Để nghiên cứu thói quen mua hàng của các khách hàng về hàng điện tử, có thể khai phá các luật kết hợp của dạng $P(X: \text{khách hàng}, W) \wedge Q(X; Y) \Rightarrow \text{mua}(X, Z)$. Điều kiện X là một khóa của quan hệ *khách hàng*, P và Q là biến vị từ, là các biến có thể được khởi tạo bởi các thuộc tính thích hợp hoặc các chiều, như một phần của chức năng dữ liệu. Các W, Y, Z là các biến đối tượng, tức các biến có thể lấy giá trị của các thuộc tính tương ứng cho khách hàng X .

¹ TV: Television: vô tuyến truyền hình.

² DVD: Digital Versatile Disc, Digital Video Disk: đĩa video.

Việc tìm kiếm các luật kết hợp bị ràng buộc bởi các siêu luật, là Tuổi (X, "30 - ") $\wedge$ thu nhập (X, "40 - 50K") $\Rightarrow$ mua (X, "V CR"¹, [2.2%, 60%]; và Nghề (X, "sinh viên") $\wedge$ age (X, "20 - 29")) $\Rightarrow$ mua (X, "máy tính") [1.4%, 70%]. Theo luật đầu thì các khách hàng ở độ tuổi 30 với mức lương hàng năm là từ 40,000 tới 50,000, với 60% sẽ mua VCR¹ và các trường hợp này chiếm vào khoảng 2.2% trong tổng số giao dịch. Theo luật thứ hai thì khách hàng là sinh viên và ở độ tuổi 20 là thích mua máy tính vào khoảng 70% và chiếm 1.4% trong tổng số giao dịch.

Đặc điểm của các thuộc tính hoặc các chiều thích hợp có thể là một chức năng phức tạp đối với người dùng ở bước đầu. Họ có thể có ý tưởng thô về các thuộc tính thích hợp, khi thăm dò khám phá. Sau đó, khi thấy rõ dữ liệu khai phá, người ta có thể bỏ bớt hay thêm vào các dữ liệu khác, liên quan mạnh về ngữ nghĩa dữ liệu cần tìm kiếm. Đó là nội dung về tìm kiếm thông tin IR. Chẳng hạn việc bán các mặt hàng liên quan chặt chẽ tới các sự kiện cụ thể như là ngày lễ Noel hoặc tới các nhóm người cụ thể; các nhân tố này không có trong yêu cầu phân tích dữ liệu. Với những trường hợp như vậy, có kĩ thuật cho phép đặc tả tỉ mỉ các dữ liệu liên quan. Các chức năng này cho phép *đánh giá* và *xếp hạng* các thuộc tính theo sự phù hợp của nó. Ngoài ra, người ta sử dụng kĩ thuật tìm kiếm các thuộc tính có tương quan mạnh về ngữ nghĩa, để tăng bộ dữ liệu khởi tạo mà người dùng đưa ra.

1.4.4. Phân cấp khái niệm

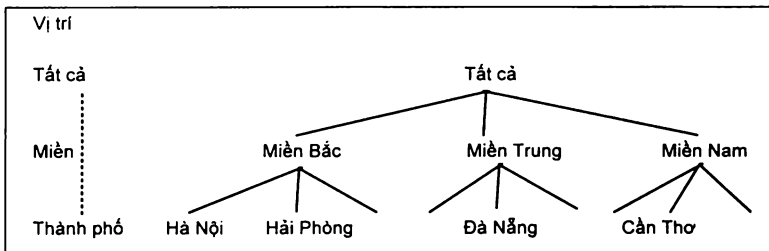
1.4.4.1. Xác định phân cấp khái niệm

Phần này trình bày về tri thức đơn giản, nhưng có khả năng mạnh; đó là phân cấp khái niệm. Phân cấp khái niệm cho phép khai phá tri thức ở nhiều mức độ trừu tượng. Nó cũng như tri thức nền.

¹ VCR Videocassette Recorder: đầu đọc video.

Định nghĩa: Tri thức nền là thông tin về lĩnh vực được khai phá mà có ích trong quá trình khai phá.

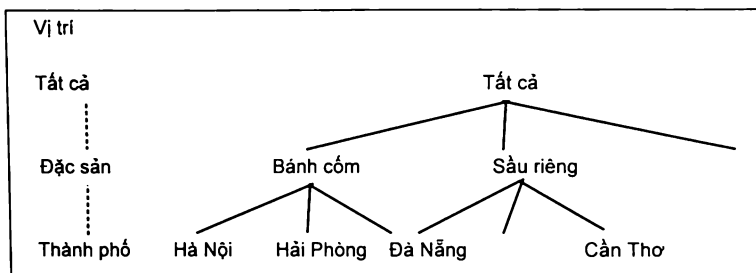
Một phân cấp khái niệm xác định luồng các ánh xạ từ một tập các khái niệm ở mức độ thấp tới mức độ cao, các khái niệm chung hơn. Một phân cấp khái niệm đối với chiều vị trí sẽ ánh xạ khái niệm mức độ thấp, như *thành phố*, tới những khái niệm tổng quát hơn nữa, như *quốc gia*. Phân cấp khái niệm này được thể hiện như tập các nút tổ chức theo dạng cây; mỗi nút thể hiện một khái niệm. Có nút đặc biệt; là gốc cây, cho biết giá trị tổng quát nhất của chiều được đưa ra. Nếu hiển thị rõ ràng thì nó đã bị ẩn. Phân cấp khái niệm này bao gồm nhiều mức độ. Theo quy ước, các mức độ trong phân cấp khái niệm này được đánh số từ cao xuống thấp bắt đầu với mức độ ứng với *tất cả*. Trong thí dụ, mức độ 1 thể hiện khái niệm *quốc gia*; mức độ 2 thể hiện các khái niệm *thành phố*. Các lá của hệ thống phân cấp tương ứng với các giá trị dữ liệu thô của chiều, tức cấp dữ liệu nguyên thủy. Một phân cấp khái niệm thường xác định sự phân lớp, thể hiện dưới dạng cây; nó cũng có thể có dạng như *lưới*.



Hình 1.20. Phân cấp khái niệm đối với chiều vị trí

Các phân cấp khái niệm là một dạng có ích của kiến thức nền, cho phép dữ liệu thô được xử lý ở mức độ trừu tượng cao hơn, tổng quát hơn. Tổng quát về dữ liệu của dữ liệu, hoặc ngược lại, tức đặc biệt hóa, đạt được nhờ việc thay thế mức độ dữ liệu nguyên thủy. Chẳng hạn thay *thành phố* cho *vị trí*; hay giá trị số thay cho *tuổi*; hoặc sử dụng các mức độ khái quát hơn, như *vùng*

địa lý, khoảng tuổi ... Điều này cho phép người dùng quan sát thấy dữ liệu có ý nghĩa hơn và trừu tượng, khiến các mẫu khai phá dễ hiểu hơn. Nhìn chung, đó là lợi thế của việc *cô đọng, nén* dữ liệu. Khai phá trên một bộ dữ liệu nén sẽ yêu cầu các thao tác ra/vào ít hơn và hiệu quả hơn, so với việc khai phá trên bộ dữ liệu lớn và không nén.



Hình 1.21. Phân cấp khái niệm khác, đối với chiều vị trí, theo đặc sản địa phương

Nếu kết quả dữ liệu thể hiện điều chung chung, các phân cấp khái niệm cũng cho phép chuyên môn hóa hoặc *đào sâu*; theo đó các giá trị khái niệm được thay thế bởi các khái niệm mức độ thấp hơn. Nhờ thao tác *quay lại và đào sâu*¹ người dùng có thể soi di dữ liệu từ quan điểm khác nhau, hiểu biết sâu hơn về bên trong cá mối quan hệ dữ liệu ẩn. Các phân cấp khái niệm có thể thu nhận được từ người dùng hệ thống, các chuyên gia về lĩnh vực, hoặc các *kiến trúc sư tri thức*. Các phân cấp khái niệm có thể thường được khai phá tự động hoặc định nghĩa động dựa trên việc phân tích thống kê và phân phối dữ liệu. Đó là vai trò của phân tích thống kê, liên quan chặt với khai phá dữ liệu.

Có thể có nhiều phân cấp khái niệm đối với một thuộc tính và chiều, dựa trên các quan điểm người dùng khác nhau. Chẳng hạn người quản lý bán hàng của *Hàng điện tử* quan tâm nghiên cứu và thói quen mua hàng của khách hàng tại các địa điểm khác nhau

¹ Roll-up, drill-down: quay lại, đào sâu.

Phân cấp khái niệm về vị trí sẽ có ích đối với nhiệm vụ khai phá. Giả sử người quản lý tiếp thị phải nghĩ ra các chiến dịch quảng cáo cho *Hàng điện tử*; chẳng hạn lưu ý đến cả ngôn ngữ trong tiếp thị, quảng cáo...

Có bốn kiểu chính của các phân cấp khái niệm; kiểu phổ biến nhất (i) phân cấp theo lược đồ; (ii) phân cấp theo nhóm. Ngoài ra người ta cũng nghiên cứu (i) phân cấp xuất phát theo thao tác; (ii) phân cấp dựa trên luật.

1.4.4.2. Phân cấp lược đồ

Phân cấp theo lược đồ nhấn mạnh mối quan hệ ngữ nghĩa giữa các thuộc tính. Thông thường một phân cấp theo lược đồ mô tả một chiều của kho dữ liệu.

Định nghĩa: Phân cấp lược đồ là toàn bộ hay một phần có thứ tự giữa các thuộc tính trong lược đồ cơ sở dữ liệu.

Lược đồ quan hệ cho *địa chỉ* bao gồm các thuộc tính *phố*, *thành phố*, *quốc gia*. Người ta có thể định nghĩa phân cấp lược đồ *vị trí* bởi thứ tự *phố* < *thành phố* hay *thành phố* < *quốc gia*. Điều này có nghĩa là *phố* có mức khái niệm thấp hơn *thành phố*, *thành phố* thấp hơn *quốc gia*. Phân cấp theo lược đồ cung cấp thông tin, tức siêu dữ liệu, tức dữ liệu về dữ liệu. Việc xác định dữ liệu phải (i) liên quan tới giao tác người dùng; (ii) đặc tả các thuộc tính thích hợp cho khai phá dữ liệu. Chẳng hạn, khi xác định thuộc tính liên quan tới bất kỳ phân cấp theo lược đồ chứa *thành phố*, người ta có thể tự động đưa ra các thuộc tính liên quan; tuy nhiên thuộc tính đó không cần thiết cho quá trình khai phá dữ liệu về sau.

1.4.4.3. Phân cấp theo nhóm

Toàn bộ hoặc một phần dữ liệu có thể được định nghĩa theo nhóm. Phân cấp theo nhóm có thể dùng để cải tiến hoặc làm phong phú hơn phân cấp theo lược đồ khi hai kiểu này kết hợp

với nhau. Chúng thường được sử dụng để xác định một tập nhỏ các quan hệ giữa các đối tượng.

Định nghĩa: Phân cấp theo nhóm là tổ chức các giá trị cho thuộc tính hoặc chiều đưa ra thành nhóm của các hằng số hoặc phạm vi giá trị.

Một phân cấp theo nhóm của các thuộc tính tuổi có thể được đặt tả trong phạm vi $\{20 - 39\} \subset \text{trẻ}$; $\{\text{trẻ, trung niên, già}\} \subset \text{tất cả (tuổi)}$. Đặc tả phạm vi giống nhau tương tự.

1.4.4.4. Phân cấp theo thao tác

Các thao tác phân cấp theo thao tác được xem là (i) giải mã thông tin từ xâu mã hóa; (ii) trích dẫn thông tin từ các đối tượng dữ liệu phức tạp; (iii) phân cụm dữ liệu.

Định nghĩa: Phân cấp theo thao tác là phân cấp dựa trên các thao tác được đặc tả bởi người dùng, các chuyên gia, hoặc hệ thống khai phá dữ liệu.

Chẳng hạn đối với thao tác duyệt, địa chỉ email hoặc một URL của WWW chứa các thông tin phân cấp sở, ngành, các trường đại học và các nước. Các thao tác giải mã có thể được định nghĩa để trích xuất thông tin để tạo thành dạng phân cấp khái niệm. Chẳng hạn địa chỉ email tuandt@vnu.edu.vn thể hiện theo thứ tự *tên người dùng, đơn vị công tác, phân lớp ngành kinh tế, tên nước*; nó sử dụng cho phép duyệt trên WWW.

1.4.4.5. Phân cấp theo luật

Tri thức luật, tức luật IF THEN, là loại tri thức cấu trúc, thông dụng đối với các ứng dụng trong thế kỉ XX.

Định nghĩa: Phân cấp theo luật là phân cấp dựa trên luật cháy, khi toàn bộ phân cấp khái niệm hoặc một phần của nó được định nghĩa bởi một tập các luật, và được đánh giá động dựa trên cơ sở dữ liệu hiện tại và việc định nghĩa các luật.

Các luật sau có thể dùng để phân lớp các mặt hàng *Hàng điện tử* như *Giá thấp (X)*: - $giá(X, P_1) \wedge giá(X, P_2) \wedge (P_1 - P_2 < 50)$ ¹. Mặt hàng có số dư lợi nhuận ít hơn 50 được định nghĩa là mặt hàng có giá thấp...

1.4.5. Độ đo về tiêu chí quan tâm

Liên quan đến mẫu thống kê, mặc dù đặc điểm của dữ liệu liên quan tới giao tác và các loại tri thức được khai phá, có thể làm giảm đáng kể số lượng *mẫu*² được sinh ra. Quá trình khai phá dữ liệu vẫn sinh ra một số lượng lớn mẫu. Thông thường chỉ một phần nhỏ của các mẫu này là mẫu mà người dùng thực sự quan tâm. Vì vậy người dùng cần tiếp tục giới hạn số lượng mẫu không được quan tâm, mà quá trình trả về. Như vậy, có thể xác định các phương pháp cần thiết để ước lượng (i) tính đơn giản; (ii) tính tiện dụng; (iii) độ chắc chắn; và (iv) tính mới của các mẫu.

Người ta nghiên cứu (i) một vài phương pháp về lợi nhuận của mẫu; (ii) phương pháp dựa trên cấu trúc của mẫu và thống kê chúng. Nhìn chung mỗi phương pháp gắn với một *ngưỡng*, là cái mà người dùng có thể được kiểm soát được. Các luật không phù hợp với ngưỡng được xem như là không hấp dẫn; vì vậy nó không được hiện ra cho người dùng như là tri thức.

1.4.5.1. Tính đơn giản

Một nhân tố cần thiết đối với mẫu là đơn giản, phù hợp với nhận thức của con người. Mục tiêu đơn giản của các phương pháp lấy mẫu được xem như là chức năng về (i) cấu trúc mẫu; (ii) kích thước của mẫu dạng bit hoặc số lượng thuộc tính hay thao tác xuất hiện trong mẫu.

Độ dài của luật cũng liên quan đến tính đơn giản.

¹ Kí pháp “:-” được xem là “có nghĩa là”, viết theo biểu thức Horn.

² Sample: mẫu (dùng trong thống kê).

Định nghĩa: Độ dài luật, đối với các luật đã thể hiện dưới dạng liên hệ thông thường được định nghĩa như là số lượng các liên kết trong luật.

Thuật ngữ luật nhằm vào (i) luật IF THEN; (ii) luật kết hợp. Các luật kết hợp, nhận thức, phân lớp có chiều dài vượt người do người dùng định nghĩa, được xem như là không hấp dẫn. Với các mẫu được thể hiện dưới dạng cây quyết định, tính dễ giải có thể là hàm của số lượng lá hoặc số nút của cây.

1.4.5.2. Tính khả năng

Mỗi mẫu được khám phá có một phương pháp chắc chắn, liên quan tới sự đánh giá tính hợp lệ hoặc độ tin cậy của mẫu. Một phương pháp chắc chắn cho phép kết hợp các luật với dạng $A =$ là tin cậy. Với bộ dữ liệu liên quan tới giao tác, hoặc các giao dịch trong một cơ sở dữ liệu giao dịch, độ tin cậy của $A \Rightarrow B$ được định nghĩa là *tin tưởng* $(A \Rightarrow B) = P(B|A) =$ số bộ chứa cả A lẫn B/số chứa A.

Giả sử rằng một bộ dữ liệu liên quan tới giao tác bao gồm giao dịch từ phòng máy tính của *Hàng điện tử*. Một tin cậy của 8 đối với luật kết hợp mua (X, máy tính) $\Rightarrow$ mua (X, phần mềm) nghĩa là 85% các khách hàng mua một máy tính cũng sẽ mua phần mềm. Một giá trị tin cậy của 100% hoặc 1.0 chỉ ra rằng luật là chính xác trên dữ liệu được phân tích. Những luật như vậy được gọi là *chính xác*.

1.4.5.3. Tính tiện dụng

Công dụng tiềm năng của một mẫu là một nhân tố xác định sự hấp dẫn của nó. Nó có thể được đo bằng chức năng tiện ích như là *trợ giúp*¹. Trợ giúp của một mẫu kết hợp liên quan tới phần trăm của dữ liệu liên quan tới giao tác, cho mẫu là đúng. Với các luật quan hệ của dạng $A \Rightarrow B$, nó được định nghĩa là: *giúp* $(A \Rightarrow B) = P(A \cup B) =$ số bộ chứa cả A và B/ tổng số các bộ.

¹ Support: (đề nghị là) trợ giúp, (thay cho hỗ trợ).

Giả sử rằng một bộ dữ liệu liên quan tới giao tác bao gồm các giao dịch từ phòng máy tính của *Hàng điện tử*. Một trợ giúp 30% cho luật kết hợp nghĩa là 30% của tất cả khách hàng trong phòng máy tính mua cả máy tính và phần mềm.

1.4.5.4. Tính mới

Mẫu mới là những đóng góp thông tin mới hoặc tăng hiệu suất cho tập mẫu đưa ra. Chẳng hạn trường hợp ngoại lệ, dữ liệu có thể được coi là mới, nếu nó khác biệt theo mô hình thống kê hoặc kinh nghiệm của người dùng. Vấn đề khác đối với phát hiện tính mới là *loại bỏ các mẫu thừa*. Nếu phát hiện ra một luật là suy diễn từ luật khác, đã có sẵn trong cơ sở tri thức hoặc trong tập các luật xuất phát, thì cần nghiên cứu để loại bỏ sự thừa tiềm năng.

Khai phá với các phân lớp khái niệm có thể tạo nên lượng lớn các luật thừa. Chẳng hạn giả sử rằng các luật sau được khai phá từ cơ sở dữ liệu *Hàng điện tử*, theo phân lớp khái niệm, đối với vị trí: R1. vị trí (X , Việt Nam) $\Rightarrow$ mua (X , TV); và R2. vị trí (X , Hà Nội) $\Rightarrow$ mua (X , TV). Thí dụ minh họa vai trò của độ trợ giúp của luật. Người ta có thể quy định các ngưỡng trợ giúp tối thiểu đối với các luật có chứa các khái niệm mức thấp là *thấp hơn* các luật cùng khái niệm ở mức độ cao hơn. Ở đây người ta sử dụng luật R1; R2 có độ trợ giúp thấp hơn.

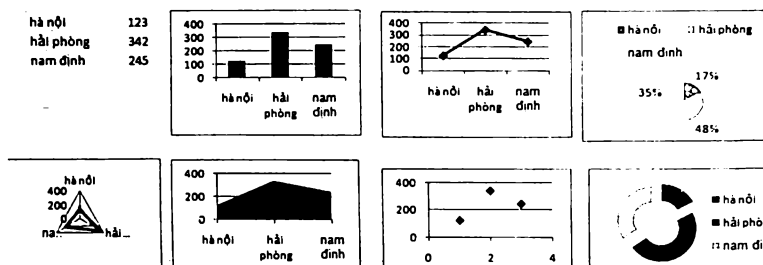
Trong khai phá dữ liệu, người dùng sử dụng dữ liệu theo phương pháp linh hoạt, tương tác cụ thể, kiểm nghiệm và thiết lập ngưỡng tương đương của chúng. Có nhiều biện pháp khách quan, ngoài những nghiên cứu cơ bản ở trên. Các biện pháp chủ quan đang sử dụng là tốt, nhưng cũng cần để cập quan hệ trong dữ liệu, ngoài các biện pháp thống kê khách quan.

1.4.6. Thể hiện và trực quan hóa các mẫu được phát hiện

Đối với khai phá dữ liệu hiệu quả, hệ thống khai phá dữ liệu sẽ có thể hiển thị các mẫu phát hiện ở nhiều hình thức, chẳng hạn

như các luật, bảng biểu, bảng chéo, bánh hoặc các biểu đồ, c quyết định, lập phương, hoặc các thể hiện trực quan khác.

Với dữ liệu hiển thị, và với mô hình phát hiện dưới nh hình thức, người dùng có thể xác định mẫu mong muốn và tươ tác hoặc hướng dẫn các hệ thống phát hiện thêm. Người sử dụ có thể xác định các loại trình bày sẽ được sử dụng để hiển thị c mẫu phát hiện.



Hình 1.22. Các dạng biểu đồ trực quan

Việc sử dụng phân cấp khái niệm đóng vai trò quan trọng trong việc giúp người dùng có hình ảnh về các mẫu phát hiện. Khai phá với phân cấp khái niệm cho phép thể hiện kiến thức phát hiện trong các khái niệm mức cao, có thể dễ hiểu hơn cho người dùng hơn so với các luật thể hiện trong điều kiện của dữ liệu nguyên thủy, dữ liệu thô, chẳng hạn như phụ thuộc hàm hay phụ thuộc đa trị, hoặc ràng buộc toàn vẹn.

Hơn nữa, hệ thống khai thác dữ liệu nên sử dụng phân cấp khái niệm, khi thực hiện (i) đào sâu; (ii) quay lại, để người dùng có thể kiểm tra phát hiện các mô hình ở nhiều cấp độ tương tự. Ngoài ra, xoay, cắt lát, đào sâu... cũng giúp người dùng xem dữ liệu tổng quát và kiến thức từ những quan điểm riêng. Một hệ thống khai thác dữ liệu cần cung cấp các hoạt động như tương tác với kích thước bất kỳ, cũng như cho các giá trị riêng của mỗi chiều.



Hình 1.23. Khả năng biểu hiện của logo của công ty IBM

Một số dạng thể hiện có thể được xem là (i) tốt hơn; (ii) phù hợp hơn, đối với các loại tri thức. Chẳng hạn (i) quan hệ tổng quát và bảng chéo tương ứng hoặc hình khoanh, biểu đồ là tốt cho trình bày các mô tả đặc tính; (ii) cây quyết định là một lựa chọn phổ biến cho các phân lớp. Các biện pháp theo sở thích sẽ được hiển thị cho mỗi mô hình phát hiện ra, để giúp người dùng xác định các mẫu cho các kiến thức có ích. Chúng bao gồm độ tin cậy, trợ giúp, và số đếm, đã được trình bày trên.

1.5. Ngôn ngữ hỏi dữ liệu

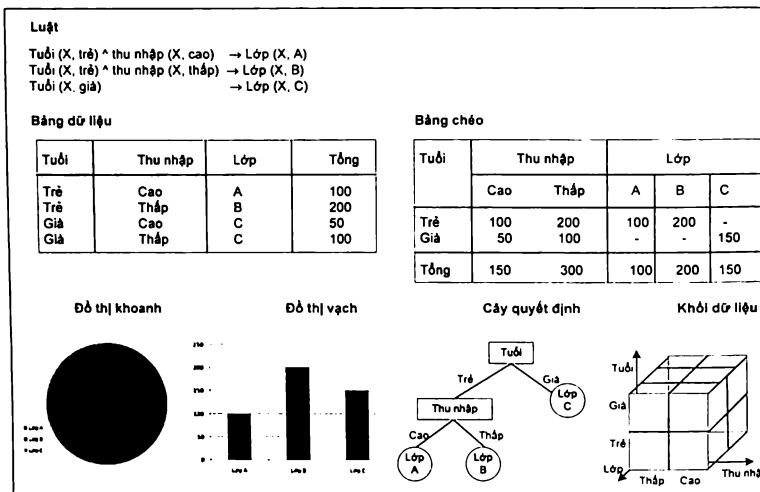
1.5.1. Vai trò của ngôn ngữ hỏi trong khai phá dữ liệu

Người ta thấy ngôn ngữ hỏi dữ liệu có vai trò quan trọng trong hệ thống khai phá dữ liệu. Đó là bởi một số lý do (i) mong muốn của hệ thống khai phá dữ liệu là khả năng trợ giúp đi kèm tương tác khai phá dữ liệu trong thứ tự thuận tiện để khám phá kiến thức một cách mềm dẻo và hiệu quả; (ii) ngôn ngữ khai phá dữ liệu có thể được thiết kế để trợ giúp tính năng trên; (iii) tương tác, để khai phá.

Định nghĩa: Ngôn ngữ hỏi dữ liệu là ngôn ngữ cho phép đặt câu hỏi để tìm kiếm, xử lý dữ liệu, tức thêm, bớt, sửa dữ liệu.

Tầm quan trọng của việc thiết kế một ngôn ngữ hỏi khai phá dữ liệu tốt cũng có thể được nhìn từ lịch sử của hệ thống cơ sở dữ liệu quan hệ. Hệ thống cơ sở dữ liệu quan hệ đã có trên thị trường trong nhiều thập niên. Các chuẩn về ngôn ngữ hỏi quan hệ, diễn ra ở giai đoạn đầu của phát triển cơ sở dữ liệu quan hệ, được xem là thành công trong lĩnh vực cơ sở dữ liệu quan hệ. Mặc dù mỗi hệ

thống cơ sở dữ liệu quan hệ thương mại có giao diện đồ họa ngu dùng của nó, lõi bên trong của mỗi giao diện là ngôn ngữ hỏi qu hệ chuẩn.

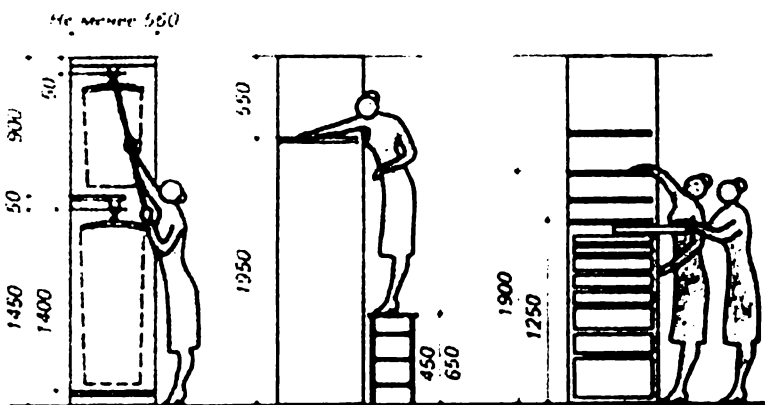


Hình 1.24. Các hình khác nhau để thể hiện và hình dung hình mẫu khám p

Liên quan đến chuẩn ngôn ngữ hỏi, các chuẩn của ngôn n hỏi quan hệ cung cấp nền tảng cho phép hệ thống quan hệ đư phát triển và tiến hóa. Nó tạo điều kiện trao đổi thông tin chuyển giao công nghệ, và thúc đẩy thương mại hóa và chấp nh rộng rãi của công nghệ cơ sở dữ liệu quan hệ. Các hoạt đ chuẩn hóa gần đây trong các hệ thống cơ sở dữ liệu, chẳng h như công cụ liên quan đến ngôn ngữ SQL, OMG, và ODMG, m họa thêm tầm quan trọng của việc có một ngôn ngữ cơ sở dữ li chuẩn, đáp ứng phát triển và thương mại hóa các hệ thống cơ sở dữ liệu. Do đó, có một ngôn ngữ hỏi tốt cho khai phá dữ liệu thể giúp chuẩn hóa sự phát triển của các nền tảng cho hệ thê khai phá dữ liệu.

Thiết kế một ngôn ngữ khai phá dữ liệu thông minh là thê thức bởi vì khai phá dữ liệu bao gồm một phạm vi rộng các nhữ vụ, từ tính chất dữ liệu tới nguyên tắc liên kết khai phá, phân l

dữ liệu, và phân tích quá trình biến đổi. Mỗi công việc có yêu cầu khác nhau. Thiết kế của một ngôn ngữ hỏi khai phá dữ liệu hiệu quả đòi hỏi một sự hiểu biết sâu sắc về sức mạnh, giới hạn, và cơ chế bản của các loại nhiệm vụ khai phá dữ liệu.



Hình 1.25. Vai trò của chuẩn thiết kế, chẳng hạn đối với độ cao

Liên quan đến quy trình thiết kế giao diện người dùng trong khai phá dữ liệu, cần có các chức năng sơ khởi, gồm:

- Tập hợp các dữ liệu liên quan đến nhiệm vụ, được khai phá;
- Các loại tri thức được khai phá;
- Tri thức nền sẽ dùng trong quá trình khám phá;
- Các biện pháp xác định ngưỡng để đánh giá mô hình; và
- Trình bày mẫu dự kiến, và phát hiện mẫu.

Dựa trên những sơ khởi, người ta thiết kế một ngôn ngữ hỏi cho khai phá dữ liệu được gọi là DMQL¹. Nó cho phép thực hiện các khai phá phụ với nhiều loại tri thức từ cơ sở dữ liệu quan hệ và dữ liệu khác, dưới nhiều cấp độ trừu tượng. Ngôn ngữ này thông qua một cú pháp giống SQL, để nó có thể dễ dàng được tích

¹ DMQL: Datamining Query Language: ngôn ngữ hỏi khai phá dữ liệu.

hợp với ngôn ngữ hỏi quan hệ, SQL. Cú pháp của DMQL được định nghĩa trong một ngữ pháp BNF¹ mở rộng.

1.5.2. Về ngôn ngữ DMQL

1.5.2.1. Ngôn ngữ hỏi

Việc trình bày chi tiết cú pháp của ngôn ngữ DMQL và minh họa qua các thí dụ sẽ chi tiết, thuận tiện cho việc sử dụng nó. Tuy nhiên, ở đây chỉ là một số khía cạnh cần thiết:

- Ngôn ngữ cần có khả năng mô tả dữ liệu và các mối quan hệ giữa dữ liệu;
- Cú pháp để xác định các loại kiến thức được khai phá. Các tri thức về (i) tính chất dữ liệu; (ii) khả năng phân biệt; (iii) các liên kết; (iv) phân lớp; (v) dự đoán;
- Cú pháp đối với đặc tả phân cấp khái niệm. Liên quan đến vấn đề này có (i) xác định lược đồ phân cấp, xác định phân cấp nhóm; (iii) xác định hệ thống phân cấp xuất phát hoạt động; (iv) xác định phân cấp dựa trên nguyên tắc;
- Cú pháp đối với đặc tả độ đo lợi nhuận;
- Cú pháp thể hiện mẫu và xác định hiển thị.

1.5.2.2. Thiết kế giao diện đồ họa với DMQL

Ngôn ngữ khai phá dữ liệu cung cấp các khái niệm cần thiết cho phép người dùng giao tiếp với hệ thống khai phá dữ liệu. Tuy nhiên, người dùng thiếu kinh nghiệm có thể phải chịu đựng ngôn ngữ khai phá dữ liệu khó dùng, với cú pháp khó nhớ. Thay vào đó, người dùng mong muốn giao tiếp với hệ thống khai phá dữ liệu thông qua một giao diện đồ họa người dùng GUI. Trong công nghệ cơ sở dữ liệu quan hệ, SQL được dùng như một lối cho ngôn ngữ hỏi, khi cài đặt hệ thống khai phá dữ liệu. Nó tạo cơ sở cho việc phát triển GUI hiệu quả.

¹ BNF: Backus Normal Form, Backus - Naur Form: dạng chuẩn Backus - Naur

Một giao diện khai phá dữ liệu bao gồm các thành phần chức năng sau:

1. *Thu thập dữ liệu và hỏi về khai phá dữ liệu.* Thành phần này cho phép người dùng xác định dữ liệu liên quan tới giao tác và để tạo ra các câu hỏi khai phá dữ liệu. Nó tương tự như GUI, cho phép hỏi dữ liệu quan hệ;
2. *Thể hiện các mẫu khai phá được.* Thành phần này cho phép hiển thị của các mẫu phát hiện dưới nhiều dạng khác nhau; gồm cả bảng biểu, đồ thị, biểu đồ, đường cong, hoặc các kĩ thuật tượng hình khác;
3. *Xác định phân cấp và xử lý.* Thành phần này cho phép người dùng đặc tả hệ thống phân cấp khái niệm. Nó thực hiện tự động hay thủ công. Ngoài ra, thành phần này cho phép người dùng thay đổi phân cấp khái niệm, hoặc điều chỉnh tự động dựa trên một bộ dữ liệu được phân phối;
4. *Xử lý các sơ khởi khai phá dữ liệu.* Thành phần này cho phép điều chỉnh động của ngưỡng khai phá dữ liệu, cũng như (i) lựa chọn; (ii) hiển thị; (iii) thay đổi phân cấp khái niệm. Nó cũng có thể cho phép thay đổi các câu hỏi khai phá dữ liệu trước đó;
5. *Khai phá tương tác nhiều mức.* Nó cho phép quay lại hoặc đào sâu trên các mẫu phát hiện được;
6. *Các thông tin khác.* Các thông tin gồm (i) sách hướng dẫn trực tuyến; (ii) chỉ mục tìm kiếm; (iii) gỡ lỗi; (iv) các phương tiện tương tác đồ họa khác.

Tuy ngôn ngữ hỏi khai phá dữ liệu chưa tiến hóa để tạo thành một chuẩn để thiết kế giao diện GUI khai phá dữ liệu, nhưng nó có thể phát triển, tạo điều kiện sử dụng phần mềm khai thác dữ liệu và hệ thống truyền thông.

Một số chức năng nguyên thủy trong GUI, chẳng hạn như chỉ vào một điểm nào đó trong một đường cong hoặc biểu đồ... khó thể hiện trong ngôn ngữ DMQL, nhưng có thể phát triển một GUI dựa trên ngôn ngữ tiêu chuẩn để thay thế ngôn ngữ SQL.

1.6. Kết luận

Người ta đã nghiên cứu năm khái niệm cơ bản về đặc tả cho một nhiệm vụ khai phá dữ liệu, thông qua việc hỏi dữ liệu. Năm khái niệm cơ bản này là (i) đặc tả của các dữ liệu liên quan, tức là dữ liệu được khai phá; (ii) kiểu tri thức được khai phá, tức là điểm, mô tả; (iii) mối quan hệ; (iv) phân lớp; (v) dự đoán.

Kiến thức cơ bản, điển hình là (i) phân cấp khái niệm; (ii) phép đo độ hấp dẫn; (iii) biểu diễn tri thức; và (iv) kỹ thuật trực quan hóa được sử dụng để hiển thị các mẫu được khai phá.

Trong định nghĩa dữ liệu liên quan tới giao tác, người dùng đặc tả cơ sở dữ liệu và các bảng, kho dữ liệu và khối dữ liệu, mà có dữ liệu sẽ được khai phá. Các điều kiện để chọn và nhóm dữ liệu và các thuộc tính, hoặc các chiều, được xem xét trong suốt quá trình khai thác.

Phân cấp khái niệm cung cấp kiến thức nền có ích cho việc khai phá các mẫu trong một mức độ ngắn gọn và thuận tiện trong việc khai phá tri thức tại các mức độ trừu tượng khác nhau.

Các phép đo lợi nhuận của mẫu được đánh giá theo (i) tính đơn giản; (ii) tính chắc chắn; (iii) tính tiện dụng; (iv) tính mới của các mẫu được khai phá. Các phép đo này có thể được sử dụng để giúp giảm số lượng mẫu không hấp dẫn khi trả lại người dùng.

Người dùng có thể đặc tả dạng mong muốn về hình ảnh cho các mẫu được khai phá như là thước, bảng biểu đồ, cây quyết định khối, đồ thị hoặc báo cáo. Các thao tác *xoay*, *đào sâu* cũng nên sẵn sàng đối với các mẫu tại các mức độ trừu tượng khác nhau.

Ngôn ngữ hỏi khai phá dữ liệu có thể được thiết kế để trợ giúp đi kèm và khai phá dữ liệu. Một ngôn ngữ hỏi khai phá dữ liệu như là DMQL cung cấp các lệnh (i) đặc tả mỗi khái niệm cơ bản khai phá dữ liệu; (ii) kế thừa khái niệm. Ngôn ngữ hỏi này dựa trên SQL và có thể là dạng chuẩn trên giao diện đồ họa người dùng đối với nhiệm vụ khai phá dữ liệu.

Dữ liệu để phân tích thống kê và khai phá

Liên quan đến phân tích thống kê và khai phá dữ liệu, người ta (i) truy cập; (ii) trích rút; (iii) tích hợp và (iv) chuẩn bị bộ dữ liệu thích hợp để khai phá. Dữ liệu đầu vào cần đảm bảo về (i) khối lượng; (ii) cấu trúc; (iii) khuôn dạng để mô hình hóa thuật toán.

Các phần trong chương này mô tả cấu trúc tổng quát đối với dữ liệu để mô hình hóa, các phép làm sạch dữ liệu và cách làm sạch. Tùy theo quan điểm của lý thuyết cơ sở dữ liệu, một dữ liệu được xem như đã sạch; nhưng theo quan điểm khai phá dữ liệu, người ta còn phải giải quyết nhiều vấn đề, chẳng hạn như dữ liệu thiếu và dữ liệu nhiễu. Nhiều thuật toán khai phá dữ liệu sẽ bỏ các trường hợp thiếu dữ liệu khi chọn các biến dự đoán.

2.1. Hiểu dữ liệu

Thường người ta cần lặp đi lặp lại các hành động hiểu và chuẩn bị dữ liệu, khi tìm hiểu bộ dữ liệu và thực hiện các phép xử lý. Các hoạt động khai phá dữ liệu trong nhiều pha liên quan đến dữ liệu số và dữ liệu phạm trù. Việc mô hình hóa sẽ thuận lợi, khi dữ liệu đã chuẩn bị phù hợp.

2.1.1. Một số định nghĩa

Trước khi xem xét chi tiết về dữ liệu, sau đây là một số khái niệm ban đầu:

- *Dữ liệu nguồn* là thông tin từ nguồn bất kì, theo dạng thú bất kì;
- *Tập phân tích* là tập các mục tin từ nhiều nguồn; *thông tin* th hiện dưới một dòng, về một thực thể;
- *Bản ghi* là một dòng trong tập phân tích;
- *Thuộc tính* là một mục tin, mô tả bản ghi theo cách nào đó;
- *Biến* là thuộc tính ứng với một cột của bản ghi thực thể;
- *Biến đích* là biến trong bản ghi thực thể, được mô hình dự đoán
- *Biến dự đoán*¹ là biến trong bản ghi thực thể, là ứng viên c trong mô hình với tư cách như dự đoán của biến đích;
- *Biến số* là biến chỉ gồm con số, được xử lí như con số;
- *Biến phạm trù* là biến có kí tự; kí tự này có thể là con số nhưng được xử lí như văn bản;
- *Biến cam* là biến tạo nên cho mỗi phần tử trong danh sách nội dung của biến phạm trù;
- *Biến đại diện*² là biến tác động lên biến đích, tương tự nh biến khác trong bản ghi.

2.1.2. Khía cạnh cần quan tâm về hiểu dữ liệu

Dưới đây là vài khía cạnh cơ bản mà người ta đề cập để hiê dữ liệu và các hoạt động liên quan:

- Cách phát hiện dữ liệu cần cho mô hình. Cần *thu thập dữ liệu*;
- Cách tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau. Cần *đến dữ liệu*;
- Dạng thức mô tả dữ liệu. Cần *mô tả dữ liệu*;
- Cách làm sạch bộ dữ liệu. Cần *đánh giá dữ liệu*.

¹ Predictor variable: biến dự đoán.

² Surogate variable: biến đại diện.

2.1.3. *Khía cạnh cần quan tâm về chuẩn bị dữ liệu*

Khi chuẩn bị dữ liệu, cần quan tâm:

- Cách làm sạch dữ liệu. Cần *làm sạch dữ liệu*;
- Cách diễn giải các biến dữ liệu. Cần *chuyển hóa dữ liệu*;
- Cách quản lí vấn đề thiếu dữ liệu. Cần *tính đến dữ liệu*;
- Phân biệt các ca xử lí. Cần *đánh trọng số và san bằng dữ liệu*;
- Về dữ liệu lẻ và dữ liệu không mong muốn. Cần *lọc dữ liệu*;
- Cách quản lí dữ liệu theo thời gian, chuỗi thời gian. Cần *trừu tượng hóa dữ liệu*¹;
- Giảm khối lượng dữ liệu. Cần *đến giảm dữ liệu*; đối với các bản ghi, cần *lấy mẫu dữ liệu*; đối với các biến, cần *giảm chiều*; đối với các giá trị, cần *rời rạc hóa dữ liệu*;
- Khả năng tạo biến mới. Cần *suy diễn dữ liệu*.

2.2. Một số nhiệm vụ đơn lẻ về chuẩn bị dữ liệu

2.2.1. *Thu thập dữ liệu*

Dữ liệu được lưu trên nhiều địa điểm của tổ chức, dưới các dạng (i) bảng tính; (ii) cơ sở dữ liệu; (iii) tư liệu in; (iv) bản viết tay. Thách thức đầu tiên là xác định vị trí dữ liệu và cách thu thập thông tin. Nếu dữ liệu tập trung tại một nơi, chẳng hạn trong kho dữ liệu, người ta vẫn cần xác định cách truy cập tốt nhất. Nếu dữ liệu có các cấu trúc khác nhau, người ta có thể có ba cách chung, để truy cập dữ liệu kinh doanh:

1. Trích dữ liệu theo câu hỏi từ cơ sở dữ liệu gồm các bảng phẳng;
2. Dùng ngôn ngữ cấp cao để truy cập trực tiếp vào tệp cơ sở dữ liệu;
3. Kết nối mức thấp để truy cập trực tiếp vào cơ sở dữ liệu.

¹ Data abstraction: trừu tượng hóa dữ liệu.

Trích dữ liệu nhờ câu hỏi cơ sở dữ liệu là cách thông dụng nhất. Ngôn ngữ thông dụng nhất là SQL-99. Các câu SQL cho phép lọc các thuộc tính đầu vào, đầu ra, dựa trên số biến, biến gộp của các biến. Tiếp cận này yêu cầu đưa bản sao dữ liệu về cùng một nơi để thực hiện đồng thời; như vậy có vấn đề với cơ sở dữ liệu lớn.

Khi dùng ngôn ngữ cấp cao, người ta có thể có ngôn ngữ hình mô hình hóa MQL¹, hay ngôn ngữ hỏi khai phá dữ liệu DMQL. Tiếp cận này chưa chuẩn hóa; nên có thể sử dụng ngôn ngữ XML.

Nhiều hệ quản trị cơ sở dữ liệu cho phép kết nối cơ sở dữ liệu ODBC³ và kết nối mức thấp, với dữ liệu đã lưu theo cấu trúc dữ liệu. Có một vài công cụ khai phá dữ liệu đã tích hợp các giao diện mức thấp để truy cập trực tiếp đến dữ liệu trong cấu trúc dữ liệu. Chẳng hạn NCR Teradata trong khai phá kho dữ liệu, năm 1999 nó truy cập trực tiếp đến dữ liệu để lập báo cáo thống kê và các phép phân tích, như hồi quy *logit*. Tiếp cận mức thấp có lợi về (i) không gian và thời gian đối với cơ sở dữ liệu lớn; (ii) quản trị dữ liệu tập trung; (iii) giảm việc sinh thêm dữ liệu không cần thiết (iv) tiện cho việc quản lý dữ liệu theo điều kiện. Tuy nhiên, truy cập mức thấp làm khó khăn cho việc đảm bảo an toàn và toàn vẹn dữ liệu.

Việc khai phá trong cơ sở dữ liệu đưa các nhiệm vụ phân tích gần hơn tới dữ liệu. Việc xấp xỉ gần với dữ liệu là hiệu quả về thời gian tính toán và giá truyền thông. Kinh nghiệm cho thấy (i) nên dữ liệu không quá nhạy cảm và bộ dữ liệu là nhỏ, nên *trích rút dữ liệu*; (ii) nếu dữ liệu quan trọng hay cơ sở dữ liệu lớn, nên truy cập dữ liệu qua khả năng *khai phá* trong cơ sở dữ liệu.

2.2.2. Trích dữ liệu

Một trong những thách thức với khai phá dữ liệu là tạo cấu trúc dữ liệu kết hợp thuận tiện cho đầu vào của công cụ khai phá:

-
- ¹ MQL: Modeling Query language: ngôn ngữ hỏi mô hình hóa.
 - ² DMQL: Datamining Query Language: ngôn ngữ hỏi khai phá dữ liệu.
 - ³ ODBC: Open Database Connectivity: kết nối cơ sở dữ liệu mở.

Các công cụ khai phá đòi hỏi thể hiện các biến như các *thuộc tính*, tức các trường, trong *dòng dữ liệu*, tức bản ghi dữ liệu.

Thí dụ có hai bảng dữ liệu R1, R2, người ta cần lấy ra những dữ liệu tổng hợp như bảng R3.

R1	Tên	Địa chỉ	Thành phố	Mã điện thoại
	Mỗ	15 Láng Hạ	Hà Nội	84 24
	Mễ	60 Kim Ngưu	Tp HCM	84 28
	Mơ	10 Văn Quán	Vĩnh Phúc	84 211

Hình 2.1. Bảng dữ liệu cho biết thông tin về khách hàng

R2	Tên	Địa chỉ	Sản phẩm	Ngày bán
	Mỗ	15 Láng Hạ	Sơn mài	15/ 03/ 2017
	Mỗ	15 Láng Hạ	Lụa	18/ 01/ 2017
	Mơ	10 Văn Quán	Sơn dầu	09/09/ 2015

Hình 2.2. Bảng dữ liệu về sản xuất

R3	Tên	Địa chỉ	Thành phố	Mã điện thoại	Sản phẩm 1	Sản phẩm 2
	Mỗ	15 Láng Hạ	Hà Nội	84 24	Sơn mài	Lụa
	Mễ	60 Kim Ngưu	Tp HCM	84 28		
	Mơ	10 Văn Quán	Vĩnh Phúc	84 211	Sơn dầu	

Hình 2.3. Bảng dữ liệu tổng hợp từ bảng R1, R2

Trong kho dữ liệu có nhiều cấu trúc từ các cơ sở dữ liệu khác nhau, nên cần có thao tác (i) *trích rút*; (ii) *chuyển đổi*; và (iii) *tải dữ liệu*¹, tức ETL. Nếu cần dữ liệu phức tạp hơn, nên sử dụng nhiều công cụ ETL trên kho dữ liệu. Công cụ ETL gây tốn kém. Hầu hết các công cụ khai phá dữ liệu có chức năng trích rút và chuyển đổi dữ liệu, nhưng thiếu chức năng tải dữ liệu.

¹ ETL: Extract, Transform, Load: trích, chuyển đổi, tải (dữ liệu).

2.2.4. Mô tả dữ liệu

Việc mô tả dữ liệu gồm (i) phân tích các số đo thống kê mô tả đối với các biến đơn; (ii) đánh giá mối quan hệ giữa các cặp biến, tức phân tích hai biến; (iii) dùng kĩ thuật đồ họa, hiển thị để cho thấy mối quan hệ phức tạp giữa các biến. Các thống kê mô tả gồm:

- *Trung bình*: cho biết giá trị trung bình, khuynh hướng hướng tâm của bộ dữ liệu;
- *Độ lệch chuẩn*: cho biết phân phối của dữ liệu quanh trung bình;
- *Cực tiểu*: giá trị nhỏ nhất;
- *Cực đại*: giá trị lớn nhất;
- *Các bảng tần suất*: cho biết phân phối tần suất của các giá trị của các biến;
- *Biểu đồ*: kĩ thuật đồ họa, cho phép hiện các giá trị tần suất trong một biến.

Phân tích và đánh giá các số đo thống kê cho phép xác định cách mô hình hóa dữ liệu. Chẳng hạn một biến có trung bình thấp và độ lệch cao sẽ ít có khả năng để dự đoán đích. Phân tích giá trị cực đại, cực tiểu và trung bình cho người ta biết về các giá trị lẻ trong bộ dữ liệu.

Thí dụ về giá trị thống kê như trong hình.

N	Trung bình	Min	Max	Độ lệch chuẩn
10000	9.769700	0.00	454.0000	15.10153

Hình 2.4. Giá trị thống kê về dữ liệu

Với bộ phần mềm thống kê tốt, người ta có thể tính được các số đo thống kê. Hầu hết các bộ khai phá dữ liệu đều có khả năng thống kê mô tả.

2.2.5. Đánh giá dữ liệu

Trước khi ấn định vấn đề nào đó trong bộ dữ liệu, cần phát hiện và xác định cách quản lý chúng. Một vài vấn đề sẽ rõ ràng sau các thao tác mô tả. Kiểm tra dữ liệu cũng tương tự như kiểm toán trong kế toán, và gồm hai phép toán (i) *trắc diện*¹ dữ liệu; (ii) *phân tích tác động* của dữ liệu chất lượng thấp. Dựa trên hai phép toán này, bộ khai phá dữ liệu có thể xác định đúng vấn đề.

2.2.6. Trắc diện dữ liệu

Khi nhìn vào phân phối dữ liệu của mỗi biến, người ta thấy:

- Khuynh hướng *hướng tâm* của dữ liệu trong biến này;
- Các dữ liệu *lỗi*;
- Số lượng và *phân phối* của khoảng dữ liệu, đối với tất cả các trường hợp;
- Dữ liệu *nghi ngờ*, dữ liệu *huấn luyện*, dữ liệu *thử hệ thống*.

Những phát hiện này sẽ được thể hiện dưới dạng các báo cáo gắn với mốc thực hiện công việc.

2.2.7. Làm sạch dữ liệu

Việc làm sạch dữ liệu gồm các phép (i) chỉnh lí dữ liệu sai; (ii) lọc dữ liệu sai ra khỏi bộ dữ liệu; (iii) lọc dữ liệu quá chi tiết, để sử dụng cho quá trình mô hình hóa.

- *Kiểm tra*² mã so với danh sách các giá trị chấp nhận được. Việc đầu vào có thể sai, một vài đoạn mã có thể không thực hiện... cần được kiểm tra theo nội dung của biến trong tất cả các bản ghi, để đảm bảo nội dung được kiểm tra đối với

¹ Data profiling: trắc diện dữ liệu.

² Valid: kiểm tra, theo yêu cầu đặt ra, theo các thông số kiểm tra; Verify: kiểm tra, theo chuẩn mực, quy trình; có thể một sản phẩm đáp ứng *verify*, nhưng

mỗi biến. Nhiều công cụ khai phá dữ liệu có khả năng trực diện dữ liệu, chẳng hạn SPSS;

- *Loại bỏ các bản ghi xấu.* Người ta thường xóa các dữ liệu không phù hợp trong bộ dữ liệu, do chúng sẽ gây lộn lộn trong mô hình hóa dữ liệu, tức làm giảm năng lực dự đoán của mô hình.

2.2.8. Chuyển đổi dữ liệu

2.2.8.1. Dữ liệu số

Nhiều chương trình thống kê tham số giả thiết tác động của mỗi biến lên đích là tuyến tính. Điều này có nghĩa biến X tăng lượng a , thì biến đích sẽ tăng một số lần của a . Kiểu tăng này tạo nên hồi quy hình học. Khi hệ số tăng không cố định, việc thay đổi tạo nên mẫu lữ thừa. Nếu sử dụng thuật toán mô hình thống kê tham số, người ta sẽ chuyển hóa bất kì biến nào thể hiện đường phi tuyến; và sai số ước lượng do vi phạm giả thiết về tuyến tính sẽ không dùng để kiểm tra kết quả dự đoán của mô hình.

Một vài thuật toán thống kê hay học máy làm việc tốt nếu các dữ liệu số đã chuẩn hóa. Theo thống kê, *chuẩn hóa* có nghĩa chuyển đổi tất cả các giá trị số sang khoảng chung. Chiến lược chung là sử dụng giá trị z , dựa trên trung bình và độ lệch chuẩn. Mỗi giá trị trong biến được thay thế bởi giá trị z , hay lệch chuẩn, theo $z = (\text{giá trị} - \text{trung bình}) / \text{lệch chuẩn}$. Khoảng giá trị z là $(-\infty, \infty)$, nhưng phân phối chuẩn, tức 99.75% các giá trị sẽ thuộc vào $(-3; 3)$. Hầu hết bộ phần mềm thống kê hay khai phá dữ liệu đều sử dụng giá trị chuẩn. Các thuật toán chuẩn hóa khác giới hạn các giá trị vào $(-1; 1)$; chẳng hạn thuật toán máy vectơ tựa SVM, mạng nơron nhân tạo ANN...

2.2.8.2. Các biến phạm trù

Các biến phạm trù có vấn đề riêng. Một vài biến phạm trù có giá trị số nguyên 1..9 sẽ được giả thiết là biến theo số liên tục

trong thuật toán mô hình hóa thống kê tham số. Các biến đó vẫn sử dụng được, nếu nhận giá trị không nguyên, chẳng hạn giá trị 1.23, chưa xác định trong bộ dữ liệu. Tuy nhiên, các biến khác chứa các phạm trù văn bản thay vì số, chẳng hạn, *mục*¹ gồm các màu xanh, vàng, trắng... cần đến các biến cầm. Biến cầm là biến nhị phân, giá trị 0, 1, phản ánh sự có mặt của mã phạm trù cụ thể trong biến đã biết.

Trường hợp	Màu	Màu đỏ	Màu xanh	Màu vàng	Màu xanh lá cây
1	Đỏ	1	0	0	0
2	Xanh	0	1	0	0
3	Vàng	0	0	1	0
4	Xanh lá cây	0	0	0	1
5	Xanh	0	1	0	0

Hình 2.5. Các biến phạm trù

Các thuật toán lệ thuộc vào tính toán hiệp phương sai, tức hồi quy, hay cần đến các phép toán số, như mạng nơron, cần làm việc với con số. Các biến cầm chuyển dữ liệu phạm trù, rời rạc, thành dữ liệu số. Việc thêm các biến cầm vào phân tích sẽ giúp tạo nên mô hình phù hợp hơn, nhưng cần tăng chi phí. Mỗi biến thô dùng đến nhóm các biến cầm sẽ khiến mất đi một bậc tự do trong phân tích. Số bậc tự do thể hiện số các mục độc lập về thông tin để ước lượng mục thông tin khác, tức biến đích. Do vậy, nếu càng đáp ứng mô hình, tức mô hình càng *chính xác*, người ta càng mất bậc tự do. Nếu có ít thông tin, người ta càng có ít khả năng áp dụng thành công mô hình này trên bộ dữ liệu khác, dù mẫu đích chỉ hơi khác với cái đã khớp với mô hình. Tính hướng này được gọi là *giảm tính tổng quát* của mô hình. Tính tổng quát là quan trọng đối với sự chuẩn xác² của mô hình.

¹ Entry: mục.

² Precision: chính xác, cho biết kết quả chụm; Accuracy: chuẩn xác, cho biết kết quả chụm và gần đích.



Hình 2.6. Chính xác và chuẩn xác

Vấn đề với các biến cam xuất hiện trong các thuật toán thống kê tham số, thuật toán học máy; chẳng hạn thuật toán cây quyết định hay mạng nơron, với khuynh hướng gây ra *quá khít*¹.

Định nghĩa: Quá khít là kết quả của việc cố khớp với giải pháp của mô hình đối với bộ dữ liệu huấn luyện, nên không thực hiện tốt đối với bộ dữ liệu khác.

Có thể huấn luyện mô hình hầu như hoàn hảo với thuật toán học máy, nhưng không hữu hiệu để dự đoán đầu ra, tức biến đích, trên bộ dữ liệu khác.

Nội dung của vài biến được mô tả dưới dạng xâu kí hiệu phạm trù, nhưng chúng thể hiện xu thế về số. Chẳng hạn, trong phân tích việc tăng sức ép về công việc, người ta có thể dùng biến đối với ngày trong tuần. Dù rằng giá trị *hai, ba, tư...* là kiểu phạm trù, nhưng chúng gắn với mức độ về sức ép công việc, nên có thể xử lí như kiểu số. Người ta có thể thể hiện chúng để tiện cho diễn giải kết quả phân tích. Các biến phạm trù khác thực sự thuộc kiểu phạm trù, không phản ánh bất kì cơ sở số nào. Đối với các biến phạm trù này, người ta tạo các biến cam đối với mỗi phạm trù và đưa chúng vào mô hình theo cách tách biệt.

2.2.9. Thay thế dữ liệu

Khi dữ liệu thiếu trong biến, đối với trường hợp cụ thể, cần thay nó bằng dữ liệu trực quan, nếu có thể. Việc bổ sung ước

lượng hợp lý đối với giá trị dữ liệu phù hợp đối với biến này là tốt hơn để trống. Phép toán quyết định giá trị nào sẽ được chọn là *thay thế dữ liệu*¹.

Định nghĩa: Thay thế dữ liệu là phép gán cho biến thiếu giá trị một giá trị phù hợp.

Trong thống kê, dữ liệu thiếu có thể gán là trống. Để xác định giá trị hợp lý, cần để ý đến (i) quy tắc đếm; (ii) ngữ cảnh.

Lựa chọn kỹ thuật phù hợp để quản lý các giá trị thiếu phụ thuộc vào giả thiết về *mẫu dữ liệu thiếu* trong bộ dữ liệu. Nếu có mẫu mạnh trong các giá trị thiếu của một biến, chẳng hạn do hỏng bộ thu nhận, biến này cần bị loại khỏi mô hình.

2.2.9.1. Giả thiết về hoàn toàn thiếu ngẫu nhiên

Giả thiết về hoàn toàn thiếu ngẫu nhiên MCAR² thỏa mãn theo định nghĩa sau.

Định nghĩa: Xảy ra thiếu dữ liệu hoàn toàn ngẫu nhiên MCAR khi xác suất xuất hiện các giá trị thiếu trong một biến là không liên quan đến giá trị của biến ấy, hay không liên quan đến các giá trị của biến khác.

Nếu MCAR đúng, các giá trị của mỗi biến được xem là mẫu ngẫu nhiên của tất cả các giá trị của biến này trong lực lượng đã có, cho phép rút ra bộ dữ liệu này. Giả thiết này sẽ vô lý, hay bị vi phạm trong một số trường hợp như khi người già từ chối liệt kê tuổi của họ so với người trẻ. Tức giả thiết có lý khi biến nào đó là khó đo và được đo chỉ với tập con của bộ dữ liệu.

2.2.9.2. Giả thiết về thiếu ngẫu nhiên

Giả thiết về thiếu ngẫu nhiên MAR³ là yếu hơn MCAR.

¹ Data imputation: thay thế dữ liệu.

² MCAR: missing completely at random: hoàn toàn thiếu ngẫu nhiên.

³ MAR: missing at random: thiếu ngẫu nhiên.

Định nghĩa: Giả thiết MAR đúng khi xác suất thiếu giá trị của một biến là không liên quan đến xác suất thiếu dữ liệu trong biến khác, nhưng có thể liên quan đến giá trị của biến này.

Chẳng hạn MAR có thể đúng nếu xác suất thiếu đầu vào liên quan đến trạng thái hôn nhân, nhưng không liên quan tới các đầu vào của trạng thái hôn nhân. Nếu MAR đúng, cơ chế gây nên thiếu dữ liệu xem như *bị bỏ qua*. Tức chỉ cần biết có MAR, không cần biết lí do.

2.2.9.3. Các kĩ thuật thay thế dữ liệu

Hầu hết phần mềm thống kê và khai phá dữ liệu có phương tiện quản lí dữ liệu thiếu. Thông thường, nó giới hạn ở việc ghi đơn giản, tức thay thế giá trị thiếu bằng giá trị nào đó. Có phần mềm công cụ thống kê, như SPSS, có nhiều khối xử lí thiếu dữ liệu hoàn thiện, đảm bảo các công cụ nhiều biến có thể điền dữ liệu thiếu. Sau đây là vài kĩ thuật.

1. *Xóa danh sách, hay trường hợp.* Điều này có nghĩa toàn bộ bản ghi bị xóa khỏi phân tích. Kĩ thuật này thường mặc định cho nhiều thuật toán thống kê hay học máy. Nó có ưu điểm sau:

- Dùng cho nhiều loại phân tích khai phá dữ liệu;
- Không yêu cầu phương pháp thống kê đặc biệt;
- An toàn khi dữ liệu xuất hiện MCAR;
- Tốt cho dữ liệu với các biến hoàn toàn độc lập;
- Ứng dụng cho bộ dữ liệu hợp với hồi quy tuyến tính và hợp với hồi quy logit hay Poisson;

Tuy nhiên bất kể sự an toàn và tiện sử dụng, kĩ thuật này có nhược điểm:

- Xóa cả thông tin trên dữ liệu không thiếu của bản ghi này, khiến nội dung thông tin chung bị giảm;

- Nếu dữ liệu thỏa mãn MAR, xóa danh sách có thể sinh ra đánh giá lệch;
- 2. *Chấp nhận giá trị trống*. Nó cho phép không xem xét đến dữ liệu trống, nhưng vẫn xét các dữ liệu có mặt trên bản ghi có dữ liệu trống. Kỹ thuật này áp dụng khi các giá trị của một biến được sử dụng để tính hiệp phương sai của biến đó. Ưu điểm của kỹ thuật này là hồi quy tuyến tính có thể được ước lượng chỉ từ trung bình mẫu và ma trận hiệp phương sai, tức liệt kê hiệp phương sai của mỗi biến. Tiếp cận này chỉ đúng khi dữ liệu là MCAR. Nếu dữ liệu chỉ là MAR, tiếp cận này có thể gây thiên lệch trong ước lượng.
- 3. *Thay thế giá trị hợp lý*. Thay thế giá trị thiếu bằng trung bình của các giá trị có mặt được gọi là *thay thế trung bình*. Nếu có thể dùng luật quyết định để cung cấp giá trị đặc biệt cho giá trị thiếu, nó có thể gần với giá trị thực. Đối với vài biến, thay khoảng trống bằng *trung bình* có thể có ý nghĩa, nhưng có trường hợp thay bằng *trung vị*¹ thì hợp lý hơn. Nên người ta có thể có nhiều giải pháp và tự xác định giải pháp phù hợp.

2.2.9.4. Thay thế khả năng tối đa

Kỹ thuật thay thế dữ liệu thiếu bằng dữ liệu có nhiều khả năng nhất cần giả sử các biến dự đoán là độc lập. Nó sử dụng hàm mô tả bản đồ mật độ xác suất, tương tự như bản đồ hình dạng, để tính khả năng giá trị thiếu, theo các trường hợp mà giá trị này có mặt. Hàm tiếp theo sẽ tối đa hóa khả năng này, tương tự như phát hiện điểm cao nhất trên bản đồ hình dạng.

2.2.9.5. Thay thế nhiều lần

Thay vì lấy một giá trị thay cho dữ liệu thiếu, có tiếp cận cho phép dữ liệu này quyết định giá trị nó cần. Tiếp cận thay thế, để

¹ Median: trung vị.

thử, nhiều lần dùng nhiều biến để dự đoán giá trị nào phù hợp với dữ liệu thiếu.

1. *Thay thế ngẫu nhiên đơn giản*. Kỹ thuật này tính hồi quy trên tất cả các giá trị có mặt trong tất cả các biến để ước lượng giá trị thiếu. Tiếp cận này có xu hướng đánh giá với sai số chuẩn thấp;
2. *Thay thế ngẫu nhiên nhiều lần*. Kỹ thuật này lặp lại nhiều lần thay thế đơn giản. Phương pháp này là tin cậy hơn, do nó xét các tham số hồi quy, chẳng hạn trung bình, như các giá trị mẫu trong phân phối dữ liệu. Cải tiến của tiếp cận này là thay thế nhiều lần, với các mẫu dữ liệu khác nhau. Giá trị trung bình thay thế được tính trên nhiều mẫu và nhiều thay thế.

Hệ thống SAS có thủ tục MI cho phép thay thế nhiều lần. Còn có SAS PROC MIANALYSE cho phép phân tích đầu ra của PROC MI, cho biết thông tin bị tổn thất do thay thế dữ liệu thiếu. Bảng sau cho thấy hướng dẫn về chọn kỹ thuật thay thế đối với biến đã biết.

Bảng 2.1. Hướng dẫn chọn kỹ thuật thay thế dữ liệu

Xóa cả trường hợp	Xóa cặp	Thay thế	Thay thế ML	Kì vọng lớn nhất	Thay thế ngẫu nhiên đơn giản	Thay thế ngẫu nhiên nhiều lần
Đơn giản nhất, dễ nhất	Bảo toàn trường hợp	Tốt khi biết luật quyết định	Không lệch khi nhiều dữ liệu	Quá trình lặp	Xu hướng ước lượng quá về tương quan	Tốt nhất đối với thuật toán phi tuyến
Bỏ trường hợp này			Ước lượng bền vững dưới điều kiện diện rộng	Sử dụng tất cả các biến khác để dự đoán giá trị thiếu		Không tốt để xác định hiệu quả tương tác

<i>Xóa cả trường hợp</i>	<i>Xóa cặp</i>	<i>Thay thế</i>	<i>Thay thế ML</i>	<i>Kì vọng lớn nhất</i>	<i>Thay thế ngẫu nhiên đơn giản</i>	<i>Thay thế ngẫu nhiên nhiều lần</i>
Chấp nhận nếu số các trường hợp là lớn và số sự kiện theo mô hình là hiếm			Dữ liệu cần là MAR	Dữ liệu cần là MAR		Cần khớp với mô hình phân tích
Hầu hết được kiểm chứng thống kê			Tốt khi dữ liệu đơn điệu	Giả thiết về phân phối chuẩn		Phù hợp nếu dữ liệu bị xóa không chấp nhận xóa theo trường hợp
An toàn đối với mọi kiểu phân tích khai phá dữ liệu			Phù hợp nếu số các trường hợp bị xóa không chấp thuận xóa theo trường hợp			
Tốt đối với bộ dữ liệu có biến hoàn toàn độc lập						

1.2.9.6. Khuyến cáo về thay thế

Một vài kinh nghiệm liên quan đến việc thay thế dữ liệu trống:

- Nếu có nhiều trường hợp, nên xóa bản ghi có dữ liệu thiếu;
- Nếu dùng hồi quy tuyến tính làm thuật toán mô hình hóa và ít dữ liệu, nên chấp nhận dữ liệu trống có lựa chọn;

- Nếu có hệ thống SAS, hãy dùng PROC MI;
- Nếu có ý niệm về giá trị đầu ra, thay dữ liệu thiếu bằng giá trị hợp lý;
- Trường hợp chung, nên dùng giá trị trung bình để thay.

2.2.10. Trọng số dữ liệu và cân đối

Các thuật toán thống kê tham số đo sự khác nhau giữa các μ đo suy diễn, như trung bình, độ lệch..., trên các giá trị nguy hiểm mà đặc tính của phân phối dữ liệu xác định. Chẳng hạn nếu giá trị trong bộ dữ liệu ra ngoài 1.96 đơn vị lệch chuẩn so với trung bình tức giá trị z , thì ngoài giá trị này có phần của dữ liệu khác, để 95%; giới hạn này được gọi là mức tin cậy 95%, hay 95% CL¹. Các thuật toán thống kê tham số, như OLS, học từ dữ liệu, nhờ lượng tất cả các trường hợp để tính độ đo, và so sánh tất cả các giá trị dữ liệu với độ đo này, cùng với bảng xác suất chuẩn, để quy định giữa hai biến có quan hệ hay không. Đôi khi người ta muốn đánh trọng số của mỗi điểm dữ liệu với dữ liệu trong biến khác để tính mối quan hệ bền vững với thực tế. Chẳng hạn giá trị đầu vào từ bộ thu nhận A có thể chuẩn xác hơn bộ thu nhận B hai lần khi đó người ta dùng trọng số 2 đối với tất cả các giá trị từ A, và trọng số 1 với B.

Các thuật toán học máy ML² học theo các cách rất đa dạng. Thay vì duyệt tất cả các trường hợp để tính độ đo *cô động*³, thuật toán ML học từng trường hợp. Chẳng hạn mạng nơron nhân tạo số ngẫu nhiên đối với mỗi biến trong lần đầu duyệt dữ liệu; trong các lần sau, trọng số được điều chỉnh như lan truyền ngược, theo tác động của nó đến các trường hợp. Không đánh trọng số theo trường hợp, do các biến trong tất cả các trường hợp có cùng tác động đến việc điều chỉnh trọng số. Tính động của công cụ ML như

¹ CL. confidence level: mức tin cậy.

² ML. machine learning: học máy.

³ Summary: cô động, tóm lược.

mạng nơron là khác với phép toán của phân tích thống kê tham số. Các phân tích thống kê tham số, hay còn được gọi là các phương pháp theo *tần suất*, chỉ xét dữ liệu một lần và điều chỉnh dựa trên số đo tính được trên dữ liệu. Các công cụ ML học gia tăng. Các trọng số của biến trong mạng nơron như mức tín hiệu giữa các nơron trong não người. Người ta học do việc truyền tín hiệu mạnh hay yếu nhiều lần, và tính thời gian giữa các nơron. Trọng số của trường hợp có thể lấy từ thuật toán mạng nơron, để thay đổi tác động của các biến trong các trường hợp hiếm, lên lời giải.

2.2.11. Lọc dữ liệu và làm trơn

Định nghĩa: Lọc dữ liệu là việc xóa các dòng, hay các trường hợp, để loại bỏ thông tin không cần thiết.

Việc lọc dữ liệu làm trong sáng tín hiệu từ các biến cần mô hình hóa. Việc loại bỏ thông tin không cần thiết sẽ giảm dữ liệu ồn dưới mức có thể gây nhiễu cho quá trình phân tích. Chẳng hạn xử lý tín hiệu ra đa; tuy nhiên tín hiệu ở đây là tín hiệu dữ liệu, và cả hai loại tín hiệu đều mô tả miền thông tin cơ bản. Tín hiệu ra đa mô tả miền cơ bản về khoảng cách và vị trí. Tín hiệu hình ảnh vệ tinh mô tả miền thị giác.

2.2.11.1. Loại bỏ dữ liệu lẻ

Loại lọc đơn giản nhất là loại bỏ dữ liệu lẻ. Đôi khi người ta muốn giữ dữ liệu lẻ, tức dữ liệu bất thường. Thực tế, một vài dữ liệu lẻ có ý nghĩa đối với mô hình hóa rủi ro tài chính, hay các sự kiện hiếm. Đối với các mô hình bình thường, việc loại các dữ liệu lẻ là hợp lý. Lý do là người ta muốn mô hình sẽ cho các kết quả bình thường. Nếu bỏ dữ liệu lẻ ra khỏi bộ dữ liệu, người ta đã bỏ đi các dữ liệu ồn, đồng nghĩa với việc làm giảm tính dự đoán của mô hình này. Người ta cũng có lý khi giữ tất cả các dữ liệu trong bộ dữ liệu, do cần tính điểm các giá trị như trong phép nhân của mô hình. Khi giữ dữ liệu lẻ, người ta phải chấp nhận dự đoán sai khoảng 5%, tức chỉ đạt yêu cầu 95%.

- Nếu có hệ thống SAS, hãy dùng PROC MI;
- Nếu có ý niệm về giá trị đầu ra, thay dữ liệu thiếu bằng giá trị hợp lý;
- Trường hợp chung, nên dùng giá trị trung bình để thay.

2.2.10. Trọng số dữ liệu và cân đối

Các thuật toán thống kê tham số đo sự khác nhau giữa các đo suy diễn, như trung bình, độ lệch..., trên các giá trị nguy hiểm mà đặc tính của phân phối dữ liệu xác định. Chẳng hạn nếu giá trị trong bộ dữ liệu ra ngoài 1.96 đơn vị lệch chuẩn so với trung bình tức giá trị z , thì ngoài giá trị này có phần của dữ liệu khác, đến 95%; giới hạn này được gọi là mức tin cậy 95%, hay 95% CL¹. Các thuật toán thống kê tham số, như OLS, học từ dữ liệu, nhờ lượng tất cả các trường hợp để tính độ đo, và so sánh tất cả các giá trị dữ liệu với độ đo này, cùng với bảng xác suất chuẩn, để quy định giữa hai biến có quan hệ hay không. Đôi khi người ta muốn đánh trọng số của mỗi điểm dữ liệu với dữ liệu trong biến khác để tính mối quan hệ bền vững với thực tế. Chẳng hạn giá trị đầu vào từ bộ thu nhận A có thể chuẩn xác hơn bộ thu nhận B hai lần khi đó người ta dùng trọng số 2 đối với tất cả các giá trị từ A, trọng số 1 với B.

Các thuật toán học máy ML² học theo các cách rất đa dạng. Thay vì duyệt tất cả các trường hợp để tính độ đo cô đọng³, thuật toán ML học từng trường hợp. Chẳng hạn mạng nơron gắn trọng số ngẫu nhiên đối với mỗi biến trong lần đầu duyệt dữ liệu; trong các lần sau, trọng số được điều chỉnh như lan truyền ngược, thì tác động của nó đến các trường hợp. Không đánh trọng số thì trường hợp, do các biến trong tất cả các trường hợp có cùng tác động đến việc điều chỉnh trọng số. Tính động của công cụ ML n

¹ CL. confidence level: mức tin cậy.

² ML. machine learning: học máy.

³ Summary: cô đọng, tóm lược.

mạng nơron là khác với phép toán của phân tích thống kê tham số. Các phân tích thống kê tham số, hay còn được gọi là các phương pháp theo *tần suất*, chỉ xét dữ liệu một lần và điều chỉnh dựa trên số đo tính được trên dữ liệu. Các công cụ ML học gia tăng. Các trọng số của biến trong mạng nơron như mức tín hiệu giữa các nơron trong não người. Người ta học do việc truyền tín hiệu mạnh hay yếu nhiều lần, và tính thời gian giữa các nơron. Trọng số của trường hợp có thể lấy từ thuật toán mạng nơron, để thay đổi tác động của các biến trong các trường hợp hiếm, lên lời giải.

2.2.11. Lọc dữ liệu và làm trơn

Định nghĩa: Lọc dữ liệu là việc xóa các dòng, hay các trường hợp, để loại bỏ thông tin không cần thiết.

Việc lọc dữ liệu làm trong sáng tín hiệu từ các biến cần mô hình hóa. Việc loại bỏ thông tin không cần thiết sẽ giảm dữ liệu ồn dưới mức có thể gây nhiễu cho quá trình phân tích. Chẳng hạn xử lý tín hiệu ra đa; tuy nhiên tín hiệu ở đây là tín hiệu dữ liệu, và cả hai loại tín hiệu đều mô tả miền thông tin cơ bản. Tín hiệu ra đa mô tả miền cơ bản về khoảng cách và vị trí. Tín hiệu hình ảnh vệ tinh mô tả miền thị giác.

2.2.11.1. Loại bỏ dữ liệu lẻ

Loại lọc đơn giản nhất là loại bỏ dữ liệu lẻ. Đôi khi người ta muốn giữ dữ liệu lẻ, tức dữ liệu bất thường. Thực tế, một vài dữ liệu lẻ có ý nghĩa đối với mô hình hóa rủi ro tài chính, hay các sự kiện hiếm. Đối với các mô hình bình thường, việc loại các dữ liệu lẻ là hợp lý. Lý do là người ta muốn mô hình sẽ cho các kết quả bình thường. Nếu bỏ dữ liệu lẻ ra khỏi bộ dữ liệu, người ta đã bỏ đi các dữ liệu ồn, đồng nghĩa với việc làm giảm tính dự đoán của mô hình này. Người ta cũng có lý khi giữ tất cả các dữ liệu trong bộ dữ liệu, do cần tính điểm các giá trị như trong phép nhân của mô hình. Khi giữ dữ liệu lẻ, người ta phải chấp nhận dự đoán sai khoảng 5%, tức chỉ đạt yêu cầu 95%.

Định nghĩa: Lề là quan sát lệch khỏi các quan sát khác, như được tạo nên từ cơ chế khác.

Người ta thấy có bốn loại thuật toán phát hiện lề (i) dựa vào độ đo khoảng cách nguy hiểm; (ii) dựa vào mật độ; (iii) dựa vào các đặc tính phản ánh; (iv) dựa vào đặc tính phân phối dữ liệu.

Cần đến việc lọc phức tạp, khi phân tích dữ liệu chuỗi thời gian. Khi đó người ta thấy tương tự như xử lý tín hiệu truyền hay tín hiệu hình ảnh.

2.2.11.2. Gộp lớn hay lựa chọn nhằm thiết lập đoạn thời gian phân tích

- *Lọc chuỗi thời gian.* Việc lọc sẽ có ích đối với mô hình hóa dữ liệu chuỗi thời gian. Việc lọc tín hiệu loại bỏ các biến động về tín hiệu, như rung pha, hoặc bỏ các đỉnh trên, hay dưới của khoảng giá trị;
- *Lọc thấp.* Việc lọc thấp loại bỏ dữ liệu dưới ngưỡng chấp nhận;
- *Lọc cao.* Việc lọc này ngược với lọc thấp, cần bỏ đi các tín hiệu trên mức cho phép.

Lí do người ta nhấn mạnh việc xử lý dữ liệu chuỗi thời gian là khi hầu hết các bộ dữ liệu cần để mô hình hóa *không phải* theo chuỗi thời gian. Các kĩ thuật xử lý tín hiệu có thể ứng dụng hiệu quả đối với bộ dữ liệu khác với dữ liệu chuỗi thời gian. Khi thấy biến dự đoán trong ngữ cảnh ảnh hưởng đến dự đoán biến đích người ta có thể xem ảnh hưởng này như tín hiệu của trạng thái hay mức độ của biến đích. Một vài biến có thể có tín hiệu mạnh tức có tính dự đoán cao. Sẽ có vài dữ liệu ồn lẫn trong giá trị biến đích, vậy nên cần lựa chọn để loại bỏ các dữ liệu ồn này, theo cách tương tự như lọc tín hiệu chuỗi thời gian.

Lọc dữ liệu thấp có thể đơn giản như loại bỏ các trường hợp với giá trị thấp hơn giá trị ngưỡng. Tác động của thao tác này là loại đi các đầu vào tầm thường đối với thuật toán mô hình hóa. Ngược lại, lọc cao dùng để loại các trường hợp trên ngưỡng. Nghệ thuật thực hành trong khai phá dữ liệu là chọn ngưỡng phù hợp.

2.2.12. Tóm lược dữ liệu

Chuẩn bị dữ liệu cho quá trình khai phá dữ liệu đòi hỏi sắp xếp bộ dữ liệu theo cách phức tạp, phù hợp. Việc trực tiếp phân tích dữ liệu chuỗi thời gian là không đơn giản; nhưng người ta có thể phân tích gián tiếp với cấu trúc dữ liệu khác. Người ta sử dụng *tóm lược*¹ dữ liệu, thay vì sử dụng dữ liệu.

Bảng 2.2. Trích dữ liệu từ hệ thống hóa đơn của công ty điện thoại

Mã khách	Tháng	Sử dụng	Phải trả	Đã trả	Cân đối
1	1	26	\$21.19	\$21.19	0
1	2	91	\$74.17	\$37.08	\$37.08
1	3	43	\$35.05	\$17.52	\$17.52
1	4	74	\$60.31	\$60.31	\$0.00
1	5	87	\$70.91	\$35.45	\$35.45
1	6	99	\$80.69	\$40.34	\$40.34
1	7	60	\$48.90	\$24.45	\$24.45
1	8	99	\$80.69	\$40.34	\$40.34
1	9	68	\$55.42	\$55.42	\$0.00
1	10	50	\$40.75	\$20.38	\$20.38
1	11	38	\$30.97	\$15.49	\$15.49
1	12	92	\$74.98	\$37.49	\$37.49
2	1	20	\$16.30	\$8.15	\$8.15
2	2	26	\$21.19	\$21.19	0
2	3	38	\$30.97	\$15.49	\$15.49
2	4	61	\$49.72	\$24.86	\$24.86
2	5	84	\$68.46	\$68.46	\$0.00
2	6	84	\$68.46	\$34.23	\$34.23
2	7	35	\$28.53	\$14.26	\$14.26
2	8	31	\$25.27	\$12.63	\$12.63
2	9	26	\$21.19	\$10.60	\$10.60

Xét bảng dữ liệu về số điện thoại. Các bản ghi tương tự như được trích từ hóa đơn của công ty điện thoại. Các bản ghi này gồm chuỗi thời gian trong 12 tháng đối với từng khách hàng. Tuy nhiên, khách hàng thứ hai chỉ có 9 bản ghi, do người này rời công ty. Trong nhiều công nghiệp, như truyền thông, thiệt hại về kinh doanh này được gọi là *tiêu hao*, hay *hụt*². Người ta mô hình hóa về sự thiếu hụt này, vào năm 1998, dựa vào các *bản ghi phân tích*³ được tạo nên từ việc *xoay*⁴ ngược dữ liệu chuỗi thời gian.

¹ Abstract: tóm lược, trừu tượng.

² Churn: phân thiếu, hụt.

³ Analytical record: bản ghi phân tích.

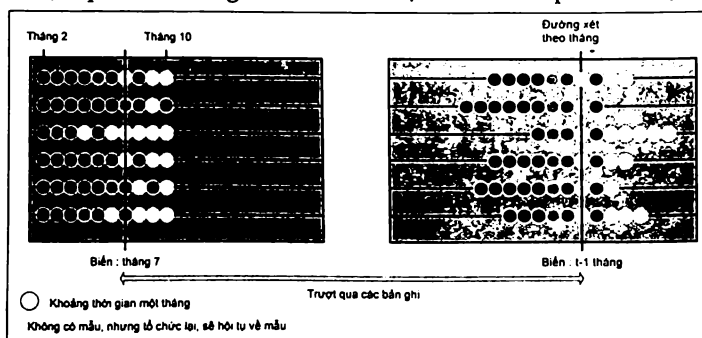
⁴ Pivot: xoay.

Bộ dữ liệu có thể được phân tích nhờ công cụ thống kê chuỗi hay học máy trên chuỗi thời gian. Người ta có thể tạo bản ghi phân tích nhờ xoay ngược: sao chép dữ liệu từ các dòng đối với khách hàng đã biết và tạo nên cột dữ liệu.

Bảng 2.3. Dữ liệu đối với 9 tháng đã được xoay

Mã khách	Sử dụng 1	Sử dụng 2	Sử dụng 3	Sử dụng 4	Sử dụng 5	Sử dụng 6	Sử dụng 7	Sử dụng 8	Sử dụng 9
1	26	91	43	74	87	99	60	99	68

Bản ghi này phù hợp với phân tích trong các công cụ thống kê và học máy. Với bản ghi này, người ta vẫn thể hiện được dữ liệu thiếu. Nhiều khách hàng bị hụt dữ liệu. Nếu sử dụng ngay các bản ghi cho quá trình phân tích mô hình, sẽ không đủ tín hiệu để thiết lập liên quan giữa dữ liệu thiếu hụt với các biến khác, vào thời điểm xác định. Người ta áp dụng kỹ thuật sử dụng trong kỹ thuật radar khi khuếch đại tín hiệu tương tự. Ở đây, người ta suy diễn tập các *rút gọn theo thời gian*¹, mà các giá trị trong mỗi biến đều liên quan với tháng bị thiếu hụt. Thay vì phân tích mối quan hệ của thiếu hụt để xác định các giá trị theo tháng, người ta liên kết tháng thiếu hụt với giá trị hóa đơn đối với tháng trước, trước nữa... Hình vẽ mô tả bậc của rút gọn thời gian để làm sáng tỏ tín hiệu hụt theo cách trực quan, và cũng rõ đối với thuật toán khai phá dữ liệu.



Hình 2.7. Mô tả bậc của rút gọn thời gian để khuếch đại tín hiệu hụt đối với hệ thống phân tích

¹ Temporal abstraction: rút gọn theo thời gian.

2.2.12.1. Rút gọn dữ liệu khác

Theo kì vọng của nhiều người khai phá dữ liệu, tự người ta rút ra một số biến dự đoán. Các rút gọn thời gian nêu trên là thí dụ về các biến suy diễn này. Cũng có thể đề ra cách rút gọn khác. Theo Lavrac, năm 2000, các rút gọn này được chia thành bốn nhóm:

1. *Rút gọn định tính*: biểu thức số ánh xạ sang biểu thức định tính. Chẳng hạn khi phân tích nhu cầu khách hàng thiếu niên so với khách hàng khác, khách hàng độ tuổi 13 đến 19 được rút gọn với giá trị 1 đối với biến *thiếu niên*; các khách hàng khác có giá trị 0;
2. *Rút gọn tổng quát*: một thể hiện xuất hiện được ánh xạ vào lớp của nó. Chẳng hạn, trong phân tích về người châu Á, so với người các châu khác, các danh sách *Việt Nam, Nhật, Trung quốc...* của biến *chủng tộc* được rút gọn về 1; ngược lại 0;
3. *Rút gọn về định nghĩa*: thể hiện mà phần tử dữ liệu từ phạm trù khái niệm được ánh xạ sang phần tương ứng trong phạm trù khái niệm khác. Chẳng hạn khi kết hợp bộ dữ liệu từ các nguồn khác nhau để phân tích nhu cầu khách hàng miền Bắc, người ta có thể ánh xạ tỉnh Nghệ An trong dữ liệu nhân khẩu sang biến khác là Khu Tư;
4. *Rút gọn thời gian*: như trên đã đề cập.

Người khai phá thường hướng đến ba dạng rút gọn đầu tiên khi tạo bản ghi. Dạng rút gọn thời gian không thông dụng. Tuy nhiên, phương pháp luận về thiết kế của vài nhà cung cấp công cụ khai phá dữ liệu có dạng rút gọn thời gian.

Nếu muốn mô hình hóa dữ liệu, mà không muốn sử dụng thuật toán chuỗi thời gian, như ARIMA, người ta có thể tạo trừu tượng thời gian và mô hình hóa chúng. Đôi khi chúng được gọi là biến *hậu quả*¹, do đáp án được mô hình hóa xuất hiện sau nguyên nhân của đáp án.

¹ Lag: kéo đi sau.

2.2.13. Rút gọn dữ liệu

Rút gọn dữ liệu gồm ba quá trình tổng quát:

1. *Giảm chiều*, tức giảm số biến;
2. *Giảm các trường hợp*, tức lấy mẫu dữ liệu;
3. *Phân biệt, rời rạc hóa* các giá trị.

2.2.14. Lấy mẫu dữ liệu

Trong khai phá dữ liệu, lấy mẫu dữ liệu nhằm bốn mục đích:

1. *Giảm số trường hợp dữ liệu tham gia vào thuật toán mô hình hóa*. Nhiều khi chỉ dùng đến 10% hay 20% dữ liệu để xây dựng mô hình dự đoán. Lấy mẫu như vậy được gọi là *lấy mẫu ngẫu nhiên đơn giản*. Cơ sở lý thuyết cho việc này là mỗi trường hợp mẫu có cơ hội như trường hợp khác;
2. *Giúp lựa chọn các trường hợp đồng nhất với phần kết quả*. Chẳng hạn mô hình hóa mẫu hành vi gọi điện thoại, người ta kiểm chứng hành vi gọi khác nhau từ vùng thành phố và nông thôn. Việc chia bộ dữ liệu như vậy được gọi là *phân đoạn cơ sở dữ liệu*; cho phép xây dựng mô hình tách biệt với mỗi vùng. Lấy mẫu như vậy được gọi là *lấy mẫu ngẫu nhiên phân tầng*; các đoạn ứng với các tầng tách biệt;
3. *Giúp cân đối tỉ lệ xuất hiện để phân tích trong công cụ học máy*. Các công cụ học máy như mạng nơron hay cây quyết định rất nhạy với bộ dữ liệu không cân đối. Dữ liệu không cân đối là loại mà một phạm trù của biến đích là hiếm so với phạm trù khác. Việc cân đối bộ dữ liệu đòi hỏi lấy mẫu phạm trù hiếm hơn trung bình, tức lấy quá lên, hay lấy mẫu các phạm trù chung ít đi;
4. *Có thể sử dụng lấy mẫu ngẫu nhiên đơn giản để chia bộ dữ liệu thành ba bộ dữ liệu để phân tích*. Các bộ dữ liệu là (i) *bộ huấn luyện*, được chọn ngẫu nhiên để huấn luyện mô hình; (ii) *bộ thử nghiệm*, dùng để đánh giá khả năng dự đoán của mô

hình, trước khi tinh chỉnh mô hình hay gia cố thêm; (iii) *bộ kiểm chứng*¹, dùng để thử hiệu quả cuối cùng của mô hình này, so với yêu cầu.

Bộ dữ liệu thử nghiệm cũng có thể dùng như bộ kiểm chứng, và bộ dữ liệu kiểm chứng dùng như bộ thử nghiệm. Cần thiết có bộ dữ liệu thử nghiệm do có thể mô hình không phù hợp với dữ liệu thử nghiệm này. Bộ kiểm chứng cần tách biệt với quá trình lập mô hình hóa.

2.2.14.1. Giảm chiều

Khi có kết quả phân tích, cần bỏ bớt biến không cần thiết. Vấn đề là quyết định biến nào là không cần. Có vài kĩ thuật cho phép xác định được biến phù hợp đưa vào thuật toán, và biến cần loại bỏ khỏi kết quả phân tích.

2.2.14.2. Hệ số tương quan

Cách đánh giá đơn giản nhất về mối quan hệ giữa các biến là tính hệ số tương quan giữa các biến. Bảng 2.4 là ma trận tương quan, cho biết các hệ số tương quan theo cặp. Theo ma trận tương quan, người ta thấy nhiều điều về bộ dữ liệu. (i) Tương quan của hầu hết các biến với tỉ lệ tội phạm là tương đối cao, nhưng với Đồng Xa thì thấp và không ý nghĩa. Điều này cho thấy Đồng Xa không là biến tốt để kết luận về mô hình. (ii) Không có tương quan nào của các biến dự đoán lớn hơn 0.9. Nếu tương quan giữa hai biến vượt 0.9, xem như ngưỡng kinh nghiệm, hoặc tác động là quá mạnh khi kết luận về mô hình. Quá nhiều biến mạnh trong mô hình sẽ không hay, có nghĩa không có giải pháp tối ưu. (iii) Quá nhiều trùng lặp trong tác động của biến mạnh, khi diễn giải kết quả.

¹ Phân biệt kiểm tra đáp ứng các điều kiện (*verification*) với kiểm chứng (*validation*) so sánh với các giá trị cụ thể.

Bảng 2.4. Hệ số tương quan đối với vài biến trong bộ dữ liệu Boston

	Tỉ lệ tội phạm	Xe bán lẻ	Đồng Xa	Cách xa trung tâm làm việc	Tỉ lệ thuế bất động sản
Tỉ lệ tội phạm	1	0.406	- 0.055	- 0.379	0.582
Xe bán lẻ	0.406	1	0.062	- 0.708	0.720
Đồng Xa	- 0.055	0.062	1	- 0.099	- 0.035
Cách xa trung tâm làm việc	- 0.379	- 0.708	- 0.099	1	- 0.534
Tỉ lệ thuế bất động sản	0.582	0.720	- 0.035	- 0.534	1

2.2.14.3. Phát hiện tương tác tự động χ^2

Thuật toán CHAID¹ thường dùng như thuật toán mô hình hóa cuối cùng. Nó có nhược điểm về hạn chế tác động khi đoán nhiều biến. Thuật toán này thông dụng để kiểm tra biến nhằm giảm chiều. Tuy nhiên nó thiên lệch về các biến với nhiều mức tách, làm sai lệch diễn giải về tầm quan trọng tương đối của biến dự đoán, khi giải thích kết quả trên biến phụ thuộc.

Mặc dù thiên lệch khi chọn biến, thuật toán này hay dùng để kiểm tra biến trong vài công cụ khai phá dữ liệu.

2.2.14.4. Phân tích thành phần chính

Thông thường, phân tích thành phần chính PCA² dùng để định tên các biến dự đoán trong bộ dữ liệu. PCA là kỹ thuật phát hiện mối quan hệ giữa các biến trong bộ dữ liệu, nhờ xác định và định lượng nhóm các *thành phần chính*. Các thành phần chính này tạo nên do kết hợp các biến đầu vào, liên quan đến biến đầu ra

¹ CHAID: Chi - square Automatic Interaction Detection: phát hiện tương tác động χ^2 .

² PCA: Principal Components Analysis: phân tích thành phần chính.

biết. Mỗi thành phần chính có vai trò giảm biến thiên trong bộ dữ liệu thô. Nên một vài thành phần chính đầu tiên diễn tả đa số cấu trúc nền của bộ dữ liệu. Các thành phần chính thường dùng để nghiên cứu như phương tiện giảm số các biến thô trong bộ dữ liệu. PCA thay các biến gốc bằng các thành phần chính đầu tiên. Tiếp cận phân tích mối quan hệ giữa các biến, cho phép tránh việc khớp nối các biến đầu vào với bất kì biến đích nào. Do vậy các thành phần chính có thể giấu sự khác nhau về lớp, liên quan đến biến đích.

Trong nghiên cứu về ứng dụng PCA đối với nhận dạng khuôn mặt, các thành phần chính nhằm diễn tả các đặc tính sai khi nhận dạng. Do vậy PCA thành công để giảm chiều, tạo tập các biến phù hợp để đưa vào thuật toán mô hình hóa.

2.2.14.5. Chỉ số Gini

Hệ số Gini¹ dùng để biểu thị độ bất bình đẳng trong phân phối thu nhập. Nó có giá trị từ 0 đến 1 và bằng tỉ số giữa phần diện tích nằm giữa đường cong Lorenz, tức đồ thị dùng để biểu diễn mức độ bất bình đẳng trong phân phối, và đường bình đẳng tuyệt đối với phần diện tích nằm dưới đường bình đẳng tuyệt đối. Giá trị Gini cực đại là 1, với nghĩa tất cả thu nhập thuộc về một nơi; cực tiểu 0 với nghĩa phân phối thu nhập theo tất cả các nơi.

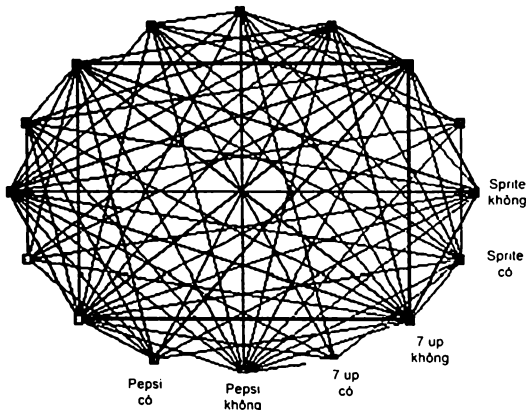
Chỉ số Gini đo độ không bằng phẳng của các giá trị trong khoảng của biến. Về lý thuyết, các biến có khối lượng không bằng phẳng tương đối lớn trong phân phối tần suất của các giá trị trong khoảng, tức giá trị chỉ số Gini cao, có xác suất cao hơn, để dùng như biến dự đoán đối với biến liên quan khác.

2.2.14.6. Các phương pháp đồ thị

Có thể dựa vào hệ số tương quan để thấy mối quan hệ giữa các biến số, nhưng không áp dụng như vậy cho các biến phạm trù

¹ Do nhà thống kê Ý, Corrado Gini đề xuất, năm 1912.

được. Một vài công cụ khai phá dữ liệu sử dụng đồ thị để xác định mức quan hệ giữa các biến phạm trù. Chẳng hạn SPSS Clementine dùng các nút Web để vạch đường giữa các biến phạm trù trên đường tròn.



Hình 2.8. Biểu đồ Web trong SPSS Clementine

Độ dày của đường liên kết cho biết độ mạnh trong mối quan hệ, chẳng hạn dựa vào biểu đồ Web để biết có mối quan hệ mạnh giữa tham chiếu *không* đối với *Pepsi ăn kiêng*, xuất hiện lúc 16.00 giờ, với *không* đối với *7Up ăn kiêng*, xuất hiện lúc 14.00 giờ.

Kĩ thuật thông dụng khác dùng để giảm chiều là:

- Lấy tỉ lệ nhiều chiều;
- Phân tích nhân tố;
- Phân rã giá trị đơn;
- Sử dụng *kĩ thuật nhân* để ánh xạ dữ liệu vào không gian chiều cao hơn. Tiếp cận này dùng trong máy vectơ tựa, học máy theo nhân...

Nhờ các phép đơn giản cũng có thể phát hiện mối quan hệ giữa các dữ liệu; chẳng hạn tính trung bình, độ lệch chuẩn, cực đại, cực tiểu và hệ số tương quan đơn giản. Có thể sử dụng điểm

Gini. Cũng có thể thu được cái nhìn toàn cảnh về mối quan hệ giữa các biến nhờ huấn luyện theo cây quyết định và suy luận chính sách nghiệp vụ.

2.2.15. Rời rạc hóa các giá trị

Vài kĩ thuật học máy có thể làm việc chỉ với các biến dự đoán phạm trù, chứ không với các biến số liên tục. Có thể chuyển các dữ liệu số sang dãy các phạm trù bằng cách gán giá trị khoảng vào một nhóm của biến mới. Chẳng hạn khoảng biến từ 1 đến 100 được phân biệt, tức chuyển sang các giá trị rời rạc, bằng cách chia thành bốn khoảng con (i) 0 - 25; (ii) 26 - 50; (iii) 51 - 75; và (iv) 76 - 100. Tên khác cho các khoảng con này là *ngăn*, *thùng*¹. Trong quá trình xếp ngăn, mỗi giá trị trong khoảng của biến được thay thế bằng số ngăn. Nhiều phân mềm khai phá dữ liệu có tiện nghi chia ngăn, cho phép tạo các khoảng con một cách tự động. Một trong những thuộc tính của quá trình xếp ngăn là giảm ồn trong dữ liệu. Việc chia ngăn là làm mềm hóa dữ liệu. Cho điểm tín dụng được tạo theo ngăn, mà biên của ngăn được điều chỉnh và lập trình để cực đại hóa năng lực dự đoán của mô hình tính điểm tín dụng.

Khi chọn tiếp cận công nghệ cho mô hình, người ta có thể nâng khái niệm công nghệ về dữ liệu đến mức nào đó trong quá trình chuẩn bị dữ liệu, nhờ (i) ghi lại dữ liệu; (ii) chuyển đổi dữ liệu; (iii) chia ngăn dữ liệu; (iv) làm mềm dữ liệu; (v) phân cụm dữ liệu.

2.2.16. Suy diễn dữ liệu

2.2.16.1. Giới thiệu hay suy diễn của biến đích

Việc giới thiệu, suy diễn biến đích xác định đích của mô hình hóa, theo các biến đầu vào. Mục tiêu của quá trình mô hình hóa xét tác động đến biến đích với dự đoán của mô hình. Thường biến đích được chọn từ các biến đã có trong bộ dữ liệu. Chẳng hạn biến đích đối với mô hình về lỗi thiết bị có thể là hiện diện hay vắng

¹ Bin: ngăn, thùng.

mặt của dữ liệu lỗi trong bản ghi dữ liệu. Trong trường hợp khác, biến đích có thể xác định theo cách phức tạp hơn. Biến đích đối với tiêu hao của khách hàng trong một mô hình xác định như lượng sử dụng điện thoại, ít nhất là 70% của kì thanh toán trước đây... Thường thì biến đích được suy diễn theo cách may rủi, có quy luật logic.

2.2.16.2. Suy diễn biến dự đoán mới

Có thể tạo biến mới từ tổ hợp các biến đã biết. Chẳng hạn từ biến kinh độ, vĩ độ, có thể sử dụng biến khoảng cách đối với vị trí trong hệ thống thông tin địa lí. Đối với biến mới, có thể thực hiện chuyển đổi dữ liệu.

2.2.16.3. Rút ra theo thuộc tính của các biến tổng quát

Năm 2006, Han và Kamber xác định kĩ thuật *quy nạp hướng thuộc tính*, nhằm tổng quát hóa danh sách các phạm trù vào một biến, tạo nên biểu thức của biến ở mức cao hơn, tổng quát hơn. Chẳng hạn khi thiếu thông tin về nghề nghiệp khách hàng, có thể tạo khái niệm về *công chức văn phòng*, dựa trên mức xác định các biến khác, như *lương, chủ, số lượng ô tô*... Biến rút ra này có thể có tính dự đoán như biến đích.

Người ta có thể rút ra các biến theo đoạn với kĩ thuật này. Chẳng hạn hỏi cơ sở dữ liệu về tiềm năng khách hàng ngân hàng so với cơ sở dữ liệu khách hàng để phát hiện gián tiếp mối quan hệ giữa tiềm năng với thông tin khách hàng.

2.3. Độ đo dữ liệu

2.3.1. Giới thiệu về độ đo

Đích của khai phá dữ liệu là phát hiện các mối quan hệ trong thế giới thực. Các lĩnh vực đa dạng của thế giới cần khám phá là khoa học, kinh doanh, khái niệm... Chúng được thể hiện qua dữ liệu.

Định nghĩa: Mục dữ liệu là thể hiện qua các giá trị của thuộc tính của thực thể.

Dữ liệu được tập hợp bằng cách khớp các thực thể trong lĩnh vực quan tâm với thể hiện kí hiệu theo nghĩa của thủ tục đo. Việc đo liên kết giá trị biến với thuộc tính đã biết của thực thể. Mỗi quan hệ giữa các đối tượng được thể hiện qua quan hệ bằng số giữa các biến. Các thể hiện số này, tức các mục dữ liệu, được lưu trữ trong bộ dữ liệu. Khai phá dữ liệu quan tâm đến các mục dữ liệu.

Định nghĩa: Khoảng cách là khái niệm cho biết việc đo giữa hai đối tượng, dựa trên vector độ đo áp trên các đối tượng.

Các hoạt động về khai phá dữ liệu không chỉ phát hiện mối quan hệ, mà còn xét về cách thức thu thập dữ liệu. Không thể có bộ dữ liệu hoàn hảo. Người ta chấp nhận dữ liệu mắc các sai sót, thiếu... do lỗi của con người hay của kĩ thuật thu thập.

2.3.2. Các kiểu đo

Việc đo có thể được chia ra vài kiểu. Việc phân biệt căn cứ vào bản chất của thuộc tính mà việc đo thể hiện, hay căn cứ vào cách sử dụng kĩ thuật đo.

- *Thang đo thứ tự*¹. Đo các đối tượng theo các mối quan hệ chính tắc, với nhiều thang đo khác nhau. Tuy nhiên thang số được quen dùng và thông dụng. Ngoài ra thang số, thang đo có thứ tự cho phép người ta dễ dàng chuyển về trật tự ban đầu;
- *Thang đo tỉ lệ*². Việc chọn tỉ lệ thang đo cho phép thể hiện đối tượng qua độ đo hợp lí, thể hiện bản chất của đối tượng. Chẳng hạn đo trọng lượng được 2, 3, 5 cho biết đối tượng thứ ba bằng tổng hai đối tượng đầu, $2 + 3 = 5$. Nếu thể hiện qua hàm mũ, dễ hiểu nhầm rằng trọng lượng đối

¹ Ordinal scale: thang đo thứ tự.

² Ratio scale: thang đo tỉ lệ.

tượng cuối bằng tích trọng lượng hai đối tượng đầu, do $x^2 * x^3 = x^5$;

- *Thang đo theo tên*¹. Các kết quả đo gắn với con số, nhưng cần xác định được tên duy nhất, với ngữ nghĩa đối với giá trị đo ấy. Có thể gán cho các con số các nhãn, tên.

Có những thang đo khác nhau. Người ta còn dùng thang đo khoảng²; giá trị đo bằng đơn vị đo nhân với hằng số, rồi cộng thêm hằng số khác. Chẳng hạn đo nhiệt độ, thời gian lịch biểu. Người ta cần đề cập khả năng đảm bảo thứ tự của các đối tượng có được bảo toàn qua phép đo.

Việc đo cho phép khớp các đối tượng, trong đó mỗi quan hệ giữa các con số trong hệ thống kinh nghiệm tương ứng với mỗi quan hệ trong hệ thống số. Do việc khớp dùng để thể hiện các mối quan hệ trong hệ thống kinh nghiệm, kiểu đo này được gọi là đo thể hiện.

Định nghĩa: Thủ tục đo thể hiện là đo theo hệ thống thực nghiệm, thể hiện giá trị đo bằng số. Các thủ tục đo xác định thuộc tính, cũng như đo thuộc tính được gọi là các thủ tục phi thể hiện.

Một cách phân biệt việc *đo thể hiện* là cách đo thể hiện liên quan đến việc hiểu điều hệ thống đang làm; *đo phi thể hiện* liên quan đến dự đoán cái hệ thống sẽ làm. Khác nhau giữa hiểu, hay mô tả, một hệ thống với dự đoán hành vi của nó sẽ được trình bày sau.

Về nguyên tắc, việc khớp được xác định theo tiếp cận đo thể hiện, tức sử dụng số ứng với tác nghiệp, người ta có thể lấy bất kì giá trị số nào. Tuy nhiên, thực tế cho thấy các dữ liệu ghi được chỉ là dữ liệu xấp xỉ theo ý niệm toán học. Có vài nhận xét:

1. Khó tránh được sai sót khi đo;

¹ Nominal scale: thang đo theo tên.

² Interval scale: thang đo khoảng.

2. Dữ liệu được ghi với số hữu hạn theo hệ thập phân, không đủ chính xác với đối tượng thực. Việc sử dụng con số xấp xỉ sẽ tác động đến quá trình phân tích.

Việc phân loại độ đo đôi khi không dựa trên các thuộc tính toán học trừu tượng của thang tỉ lệ, mà dựa vào loại kĩ thuật phân tích dữ liệu trên dữ liệu đo.

Trong ứng dụng khai phá dữ liệu, các loại thang đo được dùng nhiều là đo theo *hạng mục*¹, tức phạm trù, được dùng nhiều; nó cho phép chuyển một - một, tức theo thang đo có tên, thang đo cho phép sắp thứ tự hạng mục. Thang đo số cho phép đo với giá trị thực, hay giá trị số.

2.3.3. Đo khoảng cách

Người ta có hai cách đo mức độ giống nhau:

1. Đo *trực tiếp* đối tượng; chẳng hạn chọn mua hai cái lọ to bằng nhau;
2. Đo *gián tiếp* qua vector độ đo hay qua đặc trưng mô tả các đối tượng. Lúc này người ta có thể xác định chính xác sự *giống nhau*; vậy có thể tính độ đo tương tự hình thức.

Hai thuật ngữ *khoảng cách*, *độ đo*² thường được dùng với nhau. Thuật ngữ khoảng cách thường được dùng không hình thức để tham chiếu đến tính khác nhau theo các đặc trưng mô tả các đối tượng; chẳng hạn khoảng cách Oclit. Độ đo là số đo sự khác nhau, đáp ứng ba điều kiện:

1. $d(i, j) \geq 0$ đối với mọi i, j ; $d(i, j) = 0$ chỉ khi $i = j$;
2. $d(i, j) = d(j, i)$ với mọi i, j ;
3. $d(i, j) \leq d(i, k) + d(k, j)$ với mọi i, j, k . Điều kiện này được gọi là bất đẳng thức tam giác.

¹ Categories scale: thang đo theo phạm trù.

² Distance, Metric: khoảng cách, độ đo.

Giả sử có n đối tượng với p số đo thực cho mỗi đối tượng. Vector quan sát đối với đối tượng thứ i là $x(i) = (x_1(i), \dots, x_p(i))$, $1 \leq i \leq n$. Khoảng cách Oclit giữa đối tượng i và j được xác định

$$d_E(i, j) = \sqrt{\sum_{k=1}^p (x_k(i) - x_k(j))^2}. \text{ Việc đo này giả sử mức đo như nhau}$$

đối với các biến. Khi các biến không được đo như nhau, người ta cần có cách khắc phục sai khác về đơn vị đo. Chiến lược thường dùng là chuẩn hóa dữ liệu bằng cách chia mỗi biến cho độ lệch chuẩn của mẫu, để các biến được xem xét là quan trọng như nhau. Độ lệch chuẩn đối với biến thứ k được đánh giá

$$\sigma_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_k(i) - \mu_k)^2}; \text{ trong đó } \mu_k \text{ là trung bình đối với biến } X_k,$$

được đánh giá theo trung bình mẫu $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_k(i)$. Do vậy

$\bar{x}_k = x_k / \sigma_k$ loại bỏ tác động về thang đo.

Nếu người ta quan tâm đến sự quan trọng tương đối của các biến, có thể đánh trọng số biến đó, sau khi chuẩn hóa; trở về độ đo

$$\text{theo khoảng cách Oclit } d_{WE}(i, j) = \sqrt{\sum_{k=1}^p w_k (x_k(i) - x_k(j))^2}$$

Giả sử đo về độ cao 100 lần, độ rộng 1 lần. Kết quả đo chiều cao sẽ lấn át chiều rộng, cho dù đã sử dụng tính toán chuẩn. Về chiều cao, 99 lần đo không cho biết thêm thông tin về cái cần đo, do chúng liên quan với nhau. Để giảm sự dư thừa này, người ta sử dụng phương pháp dựa trên dữ liệu. Một tiếp cận dựa trên chuẩn hóa dữ liệu, không theo hướng của mỗi biến, như trong khoảng cách Oclit, mà tính *hiệp phương sai*¹ giữa các biến.

Hiệp phương sai mẫu giữa X và Y là $Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})$; trong đó $\bar{x}$, $\bar{y}$ là trung bình mẫu của các giá trị X , Y . Hiệp phương sai là số đo cho biết cách X , Y thay đổi cùng nhau: nó có giá trị dương lớn nếu các giá trị lớn

¹ Covariance: hiệp phương sai.

của X gắn với các giá trị lớn của Y , và nhỏ với nhỏ. Nếu các giá trị lớn của x có xu hướng gắn với giá trị nhỏ của Y , hiệp phương sai sẽ âm.

Tổng quát với p biến, người ta có ma trận hiệp phương sai $p \times p$, trong đó phần tử (k, l) là hiệp phương sai giữa biến k và biến l . Nhận xét thấy ma trận hiệp phương sai đối xứng.

Giá trị của hiệp phương sai phụ thuộc vào hạng của X và Y . Có thể loại bỏ phụ thuộc này bằng cách chuẩn hóa, tức chia các giá trị của X cho độ lệch chuẩn của nó, và Y cũng vậy. Kết quả là hệ số tương quan mẫu $\rho(X, Y)$ giữa X và Y

$$\rho(X, Y) = \frac{\frac{1}{n} \sum_{i=1}^n (x(i) - \bar{x})(y(i) - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x(i) - \bar{x})^2 \cdot \frac{1}{n} \sum_{i=1}^n (y(i) - \bar{y})^2}}. \quad \text{Khi có } p \text{ biến, ma trận hệ số}$$

tương quan mẫu $p \times p$ được tạo như ma trận hiệp phương sai. Lưu ý rằng hiệp phương sai và tương quan thể hiện sự phụ thuộc giữa các biến. Nói như vậy chính xác hơn dùng thuật ngữ phương sai tuyến tính, hay tương quan tuyến tính. Xét các điểm dữ liệu có phân phối đều quanh hình tròn có tâm ở gốc mặt hai chiều X và Y . Các biến rõ ràng phụ thuộc, nhưng không theo nghĩa tuyến tính. Do vậy sự độc lập ngầm định việc thiếu tương quan, nhưng ngược lại nói chung không đúng.

Về thí dụ đo 100 lần chiều cao và 1 lần chiều rộng, người ta giảm tác động của 100 biến tương quan bằng cách đưa ma trận hiệp phương sai trong định nghĩa khoảng cách. Người ta thu được khoảng cách Mahalanobis giữa hai số đo p chiều $x(i)$ và $y(i)$

$$d_{MH}(X, Y) = \sqrt{(x(i) - x(j))^T \Sigma^{-1} (x(i) - x(j))}, \quad \text{trong đó } T \text{ thể hiện ma}$$

trận chuyển vị, Σ là ma trận hiệp phương sai, và Σ^{-1} chuẩn hóa dữ liệu so với Σ .

Một số nhận xét:

- Độ đo Oclit cũng có thể được tổng quát hóa theo một số cách. Cách rõ nhất là độ đo Minkowski $L_\lambda \sqrt[\lambda]{\sum_{i=1}^n (x_k(i) - x_k(j))^\lambda}$,

trong đó $\lambda \geq 1$;

- Khoảng cách Oclit là trường hợp đặc biệt với $\lambda = 2$. Độ đo L_1 , hay còn được gọi là độ đo khối thành phố¹ hay Manhattan, được xác định $\sum_{k=1}^p |x_k(i) - x_k(j)|$. Khi $\lambda \rightarrow \infty$ người ta có độ đo L_∞ : $\max |x_k(i) - x_k(j)|$ theo k ;
- Có một số lớn các độ đo khác để lượng hóa, nhưng vấn đề là xác định được cái phù hợp cho trường hợp cụ thể.

Đối với dữ liệu nhị phân thay đổi nhiều, người ta tính số các biến theo hai đối tượng cùng nhận các giá trị như nhau hay khác nhau. Tất cả p biến xác định cho đối tượng i và j lấy giá trị trong $\{0, 1\}$; ô $n_{i,j}$ trong bảng ứng với $i = 1, j = 1$, cho biết có $n_{i,j}$ biến với i và j đều nhận giá trị 1.

Bảng 2.5. Thông tin về việc cùng giá trị hay khác giá trị

	$j = 1$	$j = 0$
$i = 1$	$n_{1,1}$	$n_{1,0}$
$i = 0$	$n_{0,1}$	$n_{0,0}$

Với dữ liệu nhị phân, thay vì đo sự khác nhau giữa các đối tượng, người ta thường đo độ tương tự. Có lẽ việc đo tương tự rõ nhất là hệ số khớp đơn giản, được xác định $\frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}$, là tỉ

lệ các đối tượng có cùng giá trị, trong đó $p = n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}$ là tổng các biến.

Đôi khi việc đưa ô (0, 0), hay ô (1, 1), vào là không phù hợp, tùy theo nghĩa của 0 và 1. Nhận xét này đưa đến hệ số khớp khác, tên là hệ số Jaccard, được xác định $\frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}$. Hệ số Dice mở

rộng tham số này. Nếu khớp² (0, 0) là không xác đáng thì việc không khớp (0, 1) và (1, 0) sẽ nằm giữa các khớp (1, 1) và khớp (0, 0),

¹ City – block: khối thành phố.

² Match: khớp.

theo nghĩa của sự xác đáng. Do vậy người ta cho số của các khớp $(0, 1)$ và $(1, 0)$ giảm một nửa, thu được $\frac{2n_{1,1}}{2n_{1,1} + n_{1,0} + n_{0,1}}$.

- Với dữ liệu định lượng, có nhiều kiểu đo khác đối với dữ liệu nhị phân thay đổi nhiều; và vấn đề là chọn thuộc tính sẵn có nào chứ không phải độ đo;
- Đối với dữ liệu phạm trù mà các biến có quá hai hạng mục, người ta lấy 1 cho các biến mà hai đối tượng nhất trí, và 0 nếu ngược lại. Điều này thể hiện được tổng các giá trị là phần của tổng p . Nếu thấu hiểu về hạng mục, người ta có thể xác định ma trận các giá trị thể hiện các kiểu khác nhau khi không nhất trí;
- Việc đo khoảng cách hỗn hợp có thể dùng cho kiểu dữ liệu hỗn hợp, tức có cả dữ liệu nhị phân, phạm trù, định lượng.

2.3.4. Chuyển dữ liệu

Cần thiết chuyển đổi dữ liệu khi chúng không phù hợp về dạng thức, cách thể hiện mà bài toán yêu cầu. Tuy nhiên dạng dữ liệu mà mô hình yêu cầu và bản chất dữ liệu luôn có sự liên quan.

Đôi khi người ta để ý đến suy luận thống kê hình thức, trong đó quan trọng là dạng phân phối, cần thực hiện phép chuyển để các dữ liệu xấp xỉ nhất với phân phối.

Do đích chính đang xét về khai phá dữ liệu là khám phá, người ta chuẩn bị để dự tính và tìm kiếm những thứ không ngờ. Việc chuyển đổi dữ liệu có thể phát hiện được cấu trúc dữ liệu không dễ thấy trên dữ liệu gốc. Mặt khác, đi sâu hơn về khía cạnh phát hiện cấu trúc dữ liệu sẽ đề xuất cấu trúc thuận lợi cho việc chuyển. Lúc này nên có vai trò của chuyên gia lĩnh vực, để đánh giá cấu trúc khả thi.

2.3.5. Dạng dữ liệu

Các bộ dữ liệu có được từ nhiều dạng khác nhau. Các dạng này được biết như các lược đồ. Dạng đơn giản nhất là tập số đo

vector về đối tượng $o(1), \dots, o(n)$. Đối với mỗi đối tượng, có số đo của p biến $X_1, \dots, X_p$. Do vậy dữ liệu được nhìn theo ma trận n dòng và p cột. Dạng chuẩn này được tham chiếu như *ma trận dữ liệu*¹, hay *dữ liệu chuẩn*², hay *bảng*³. Quy ước này phù hợp với tiếp cận của mô hình dữ liệu quan hệ.

Thực tế có nhiều bài toán có bộ dữ liệu gồm nhiều ma trận, bảng. Bộ dữ liệu đó được gọi là dữ liệu quan hệ *nhiều chiều*⁴. Trong nhiều trường hợp, dữ liệu muốn được thể hiện ở dạng ma trận đơn. Phép toán nối các bảng được đề xuất. Nhìn chung, không bảng đơn nào có thể thu tóm tất cả thông tin từ bộ dữ liệu quan hệ nhiều chiều. Điều quan trọng là theo quan điểm của lưu trữ hiệu quả, việc làm phẳng các dữ liệu quan hệ nhiều chiều để được một bảng đơn, có thể lớn, tạo nên dư thừa dữ liệu. Dư thừa dữ liệu được minh họa nhiều trong mô hình dữ liệu quan hệ.

Định nghĩa: Dữ liệu chuỗi thời gian là dữ liệu có giá trị liên tiếp ứng với số đo gắn với thời điểm liên tiếp.

Một vài bộ dữ liệu không thích hợp với thể hiện bảng, ma trận. Điển hình là dữ liệu chuỗi thời gian, trong đó các giá trị liên tiếp ứng với số đo gắn với thời điểm liên tiếp. Người ta thể hiện chuỗi thời gian với hai biến (i) thời điểm; (ii) số đo ứng với thời điểm đó. Đây là cách truyền thống để đưa chúng vào cơ sở dữ liệu. Về thứ tự dữ liệu, người ta ghi nhận rằng không tồn tại trật tự tự nhiên; vậy nên không cố gắng đưa ra mô hình với thứ tự dữ liệu.

Định nghĩa: Một xâu⁵ là tập các kí hiệu lấy từ tập kí tự hữu hạn, xếp thành dãy.

Dãy các giá trị từ biến hạng mục là xâu. Xâu là kiểu dữ liệu đã được xếp thứ tự và đối với xâu, không nhất thiết cần tuân theo dạng ma trận chuẩn.

¹ Data matrix: ma trận dữ liệu.

² Standard data: dữ liệu chuẩn.

³ Table: bảng.

⁴ Multirelational data: dữ liệu quan hệ nhiều chiều

⁵ String: xâu.

Một kiểu dữ liệu có thứ tự nữa là *dãy sự kiện*¹. Khi có bộ hữu hạn các kí tự theo kiểu sự kiện hạng mục, dãy sự kiện là dãy các cặp {sự kiện, thời điểm xảy ra sự kiện}, cách này tương tự như với xâu, khác ở chỗ mỗi mục tin trong dãy gắn với giá trị thời gian. Kiểu dữ liệu dãy sự kiện không là kiểu dữ liệu quen thuộc. Tuy vậy, không cần hạn chế khái niệm về dãy sự kiện, đưa về dữ liệu phạm trù; có thể đưa dãy sự kiện về sự kiện lấy giá trị thực xảy ra không đồng bộ, chẳng hạn như dữ liệu về hành vi động vật.

Các dữ liệu có thứ tự có thể được trải dài theo trục một chiều, nhưng dữ liệu khác thường trong không gian nhiều chiều. Dữ liệu không gian, địa lí, hay hình ảnh được đặt trong không gian hai, ba chiều. Cần lưu ý rằng một số biến chỉ là một phần trong lược đồ dữ liệu đã xác định, chẳng hạn biến chỉ cho biết tọa độ trong không gian. Do vậy cũng dễ chuyển dữ liệu không gian, địa lí... về dữ liệu hai chiều.

Cấu trúc phân cấp là lược đồ dữ liệu phức tạp. Để thấy rõ dữ liệu có cấu trúc, người ta dùng bảng quan hệ nhiều chiều. Thể hiện dữ liệu trong một bảng không cho phép thấy hết tính phân cấp. Chẳng hạn sẽ cần cấu trúc phân cấp đặc biệt khi tiến hành trả lời một số câu hỏi điều tra bệnh tật, cần xác định trước giới tính và độ tuổi...

2.3.6. Chất lượng dữ liệu đối với phép đo riêng

Chất lượng dữ liệu ảnh hưởng đến hiệu quả của khai phá dữ liệu; như GIGO². Trong khai phá dữ liệu, có nhiều kiểu dữ liệu, lược đồ nên phức tạp cao.

Đặc trưng hóa chất lượng dữ liệu theo hai cách (i) chất lượng của bản thân dữ liệu và trường dữ liệu; (ii) chất lượng chung của thu thập dữ liệu.

¹ Event sequence: chuỗi, dãy sự kiện.

² GIGO: Garbage In, Garbage Out: vào rác, ra rác.

Định nghĩa: Thủ tục đo chính xác¹ là việc đo với sự thay đổi ít.

Dùng quá trình chính xác, đối với một đối tượng sẽ thu được cùng số đo với nhiều lần đo trong điều kiện như nhau. Chính xác ở đây không phải là số chữ số thể hiện kết quả đo.

Định nghĩa: Thủ tục đo chuẩn xác² là việc đo không chỉ ít bị thay đổi, mà còn thu được kết quả sát với giá trị người ta tính đến, hay sát với giá trị thực.

Quá trình đo có thể là chính xác nhưng không chuẩn xác. Theo thuật ngữ thống kê, sai khác giữa trung bình của các lần đo lặp lại và giá trị thực của đối tượng là *độ lệch*³ của quá trình, thủ tục đo. Đo chuẩn xác có độ lệch nhỏ và thay đổi ít.

Khái niệm giá trị thực liên quan đến khái niệm đúng. Tuy vậy ngay lần thấy đầu tiên, chưa thể xác định được số đo chuẩn xác, như đánh giá về độ cao con người, tốc độ vật thể vũ trụ.

Một khái niệm khác là độ tin cậy của quá trình đo. Độ tin cậy của phép đo cũng như độ chính xác. Độ chính xác dùng trong khoa học tự nhiên, còn độ tin cậy dùng trong khoa học xã hội.

Định nghĩa: Độ tin cậy của phép đo là độ chính xác của việc đo, dùng trong khoa học xã hội.

Phần trên đã liệt kê hai yếu tố tạo nên số đo sai (i) độ chính xác cơ bản; (ii) phân phối số đo quanh giá trị thực. Độ chính xác ứng với độ tin cậy; sự phân phối ứng với việc kiểm định, kiểm chứng⁴. Kiểm chứng là mở rộng dùng cho thủ tục đo sát với đối tượng được đo. Trong nhiều trường hợp, cần đến công nghệ phần mềm và kinh tế học.

¹ Precise: chính xác.

² Accurate: chuẩn xác.

³ Bias: lệch.

⁴ Validity: kiểm chứng.

2.3.7. Chất lượng dữ liệu đối với việc thu thập dữ liệu

Ngoài chất lượng quan sát cá nhân, người ta cần để ý đến chất lượng thu thập dữ liệu của người quan sát. Nhiều nhà thống kê và khai phá dữ liệu tiến hành suy luận trên mẫu dân cư, tức là rút ra thông tin về toàn thể dân cư, dựa trên mẫu thể hiện một phần các đối tượng. Thống kê sử dụng các tham số để tóm tắt theo mô tả về dân cư hay phân phối của đối tượng.

Các giá trị tính được từ mẫu của các đối tượng được gọi là thống kê, được dùng như đánh giá về tham số. Chẳng hạn trung bình của mẫu được dùng như đánh giá về trung bình, tức tham số trung bình, của toàn thể dân cư hay phân phối.

Các đánh giá đó chỉ có ích nếu chúng chính xác. Đánh giá từ các mẫu khác nhau có thể có nhiều thay đổi, vậy nên không tin cậy. Hoặc việc đánh giá có thể sai lệch, quá lớn hay quá nhỏ. Tuy nhiên khi tăng số mẫu, sai lệch sẽ giảm.

Chất lượng của dữ liệu được quan tâm do người ta không điều khiển cách thu thập dữ liệu. Thuật ngữ *mẫu cơ hội*, *mẫu phù hợp*¹ đôi khi được dùng để mô tả các mẫu không thực rút ra từ dân cư quan tâm.

Các dữ liệu có thể bị sai lệch theo kiểu gọi là *xoắn*², tức biến dạng, theo các dạng sau:

- Một kiểu gây dữ liệu bị vắn vẹo do người dùng qua nhiều bước lựa chọn. Chẳng hạn các bước khai mở tài khoản, chúng thực... làm người dùng không xác định đúng dữ liệu đã khai, và dữ liệu được ghi tại nhiều chứng từ, gây sai lệch;
- Dữ liệu bị sai còn do sự không khớp giữa dân cư mẫu và dân cư quan tâm. Người ta thường nghĩ mô hình hóa dân cư để giải quyết vấn đề khó; nhưng khó thực hiện;

¹ Opportunity sample, Convenience sample: mẫu phù hợp.

² Distorte: xoắn vặn, biến dạng.

- Tránh rủi ro về dữ liệu xoắn yêu cầu kiểm chứng khai phá dữ liệu, chứ không phải các bộ dữ liệu là các mẫu từ dân cư quan tâm;
- Dữ liệu xoắn có thể được coi như dữ liệu không hoàn thiện, bị mất toàn bộ hay một phần bản ghi. Đối với dữ liệu không hoàn thiện này, người ta tự hỏi *lí do thiếu dữ liệu, thông tin cần thiết có thể trong dữ liệu bị thiếu*. Việc thiếu dữ liệu sẽ khiến quá trình suy luận không hiệu quả. Khi nhằm vào suy luận trên dữ liệu, trước hết cần có đầy đủ dữ liệu.

Các quan sát ngoài lề¹ cũng thể hiện khía cạnh hoàn toàn khác về chất lượng dữ liệu. Nếu đích của khai phá dữ liệu là kiến thiết mô hình tổng thể để hiểu thêm về dữ liệu, các vấn đề bên lề có thể xảy ra tại tâm điểm của mô hình.

Khi quan sát chỉ một biến, người ta có thể phát hiện dữ liệu lẻ bằng cách vạch các điểm dữ liệu. Các điểm nằm xa dữ liệu khác sẽ được coi là ngoài lề. Khi quan sát nhiều biến, tình huống sẽ khá nhiều.

Bất kì bộ dữ liệu lớn đều có dữ liệu đáng nghi. Người ta cần chấp nhận một số hiện tượng kém chất lượng của dữ liệu. Do quá trình khai phá dữ liệu không điều khiển việc thu thập dữ liệu, nguy cơ từ dữ liệu kém trở nên quan trọng. Hunter, 1980, phát biểu về rủi ro này “dữ liệu kém chất lượng gây ô nhiễm cho suy nghĩ trong sạch và việc ra quyết định logic. Dữ liệu sai lệch và mối quan hệ rút ra từ dữ liệu ấy có thể gây hậu quả nghiêm trọng khi soạn thảo luật định, quy tắc”.

Một số nhận xét về độ đo dữ liệu:

- Phần trên xem xét dữ liệu số. Tuy nhiên các dữ liệu phi số được quan tâm. Đôi khi các định nghĩa về mục tin riêng cũng lệ thuộc vào đối tượng của việc phân tích. Trong ngữ cảnh quản lí kinh tế, nhiều loại dữ liệu chuỗi thời gian lưu trong cơ sở dữ liệu, hơn là các mục tin riêng;

¹ Outlier: lẻ.

- Các dữ liệu số và phi số đều có vai trò cơ bản trong phân tích dữ liệu. Thông thường, dữ liệu phi số, mối quan hệ... được chuyển về dạng dữ liệu số.

2.4. Hiển thị và thăm dò dữ liệu

2.4.1. Giới thiệu về hiển thị

Phần này khảo sát các phương pháp hiển thị để tìm cấu trúc trong dữ liệu. Các phương pháp hiển thị có vị trí đặc biệt trong thăm dò dữ liệu, bởi lẽ đồ họa quan trọng đối với thị giác người dùng trong việc phát hiện cấu trúc. Các phương pháp hiển thị được dùng để hiện dữ liệu theo cách nhấn mạnh vào khả năng xử lý hình dạng, hình ảnh của con người. Tiếp cận này khác với tiếp cận xây dựng mô hình hình thức để xem các dữ liệu quan sát có lộ ra trong cấu trúc tạo dữ liệu giả thuyết. Các phương pháp hiển thị là quan trọng đối với khai phá dữ liệu còn do đây là cách tốt để lướt qua dữ liệu, nhằm phát hiện các mối quan hệ. Tuy vậy, không hẳn phương pháp hiển thị không có mặt hạn chế.

Phân tích dữ liệu thăm dò được mô tả như việc tạo giả thuyết dựa trên dữ liệu¹. Người ta xem xét dữ liệu, tìm trong cấu trúc các mối quan hệ sâu giữa các trường hợp hay các biến. Quá trình này ngược với việc kiểm tra giả thuyết, mà người ta bắt đầu từ mô hình, hay giả thuyết, đề xuất rồi thực hiện các xử lý thống kê để xác định rằng thấy dữ liệu xuất hiện trong mô hình như vậy. Thuật ngữ *dựa trên dữ liệu* cho thấy chính các mẫu trong dữ liệu tạo nên giả thuyết, chứ không như trường hợp giả thuyết được tạo trên các tham số lý thuyết.

2.4.2. Thí dụ về tóm tắt dữ liệu

Đã có các khái niệm trung bình, tóm tắt đơn giản của trung bình của tập giá trị. Giả sử $x(1), \dots, x(n)$ là tập n giá trị dữ liệu.

¹ Data - driven hypothesis generation: tạo giả thuyết dựa trên dữ liệu.

Trung bình mẫu¹ được xác định $\bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$. Giá trị μ được dùng để chỉ trung bình thực của dân cư, còn $\bar{\mu}$ chỉ đánh giá dựa từ mẫu của giá trị trung bình này. Trung bình mẫu có tính hướng tâm, theo nghĩa cho biết giá trị nhỏ nhất của tổng khoảng cách bình phương giữa nó và các dữ liệu. Do vậy, nếu có n giá trị dữ liệu, trung bình là giá trị mà n lần sẽ là tổng các giá trị dữ liệu

Trung bình là số đo cho biết vị trí. Ngoài ra còn có khái niệm *nằm giữa*, cho biết giá trị có một nửa trên nó, một nửa dưới nó

Phương sai được xác định $\sigma = \sqrt{\sum_{i=1}^n (x(i) - \mu)^2}$. Do trung bình thiểu hóa tổng độ lệch bình phương, nên có liên hệ chặt chẽ giữa trung bình và phương sai. Nếu đã biết μ , trong thực tế người ta thay μ bằng $\bar{\mu}$, được đánh giá bằng dữ liệu. Khi thay như vậy, người ta có đánh giá không chệch, và phương sai là: $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mu})^2$.

lệch chuẩn là căn của phương sai, tức $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x(i) - \mu)^2}$

Độ đo vị trí khác được dùng là *phân tư đầu tiên*, có nghĩa là giá trị mà lớn hơn 1/4 của các điểm dữ liệu; và *phân tư thứ ba* lớn hơn 3/4. Trong một vài ứng dụng, khoảng giữa các phần tư² là sai khác giữa ba phần tư và một phần tư. *Khoảng*³ là sai khác giữa điểm dữ liệu lớn nhất và nhỏ nhất.

*Tính xiên, bất đối xứng*⁴ đo một phân phối có đuôi dài hay không, thường được xác định theo $\frac{\sum_{i=1}^n (x(i) - \bar{\mu})^3}{(\sqrt{\sum_{i=1}^n (x(i) - \bar{\mu})^2})^3}$.

¹ Sample mean: trung bình mẫu.

² Interquartile range: dải một phần tư.

³ Range: khoảng.

⁴ Skewness: xiên, lệch.

2.4.3. Các công cụ hiển thị các biến đơn

Một trong những cách hiển thị cơ bản nhất là biểu đồ *tần suất*¹.

Định nghĩa: Biểu đồ tần suất là biểu đồ hiện lên một số các giá trị của biến nằm trong một khoảng liên tục.

Với lượng dữ liệu không nhiều, biểu đồ này có thể làm người dùng hiểu sai về xu thế của dữ liệu. Tuy nhiên, nhược điểm này được hóa giải nhờ việc làm trơn. Cách đánh giá trơn dùng rộng rãi là đánh giá *nhân*².

Giả sử biến X có giá trị đo $\{x(1), \dots, x(n)\}$. Phân phối của điểm $x(i)$ được đánh giá tại x^* nào đó, tùy thuộc vào $x(i)$ và x^* xa nhau ra sao. Mở rộng phân phối này dựa trên dạng của hàm nhân được dùng và độ rộng của nó. Hàm nhân theo hệ số K và độ rộng h cho phép đánh giá mật độ tại điểm x là $\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x(i)}{h}\right)$, trong đó $\int K(t)dt = 1$, để đảm bảo $f(x)$ có tích phân 1, tức đậm đặc, và tại nơi hàm nhân K được chọn, hàm trơn với đỉnh là 0.

Dạng thông dụng của hàm K là đường chuẩn, hay đường Gauss, với h là tham số trải, tức độ lệch chuẩn $K(t, h) = C \cdot e^{-\frac{1}{2}(\frac{t}{h})^2}$; trong đó C là hằng số chuẩn hóa, $t = x - x(i)$ là khoảng cách từ điểm cần có đến điểm dữ liệu $x(i)$. Độ rộng h tương đương với σ , là độ lệch chuẩn, hay độ rộng, của hàm nhân Gauss.

Có các phương pháp chuẩn để tối ưu mức phù hợp của các đánh giá này với phân phối chưa biết do các dữ liệu tạo nên, nhưng ở đây chỉ xem xét khía cạnh hiển thị đồ thị.

Biểu đồ tần suất, các đường nhân, và các hiển thị một chiều khác có thể được dùng tách biệt theo nhóm. Khi có nhiều nhóm, nên dùng hiển thị theo khối, hộp. Một hộp có các búi dữ liệu; một đường đi qua hộp cho biết số đo về vị trí, thường là khoảng giữa dữ liệu.

¹ Histogram: biểu đồ.

² Kernel: nhân.

2.4.4. Các công cụ hiển thị mối quan hệ giữa hai biến

Các *điểm rải*¹ là công cụ chuẩn để hiển thị hai biến cùng lúc. Theo biểu đồ điểm rải dễ thấy tương quan giữa hai biến, khi một giá trị cao, giá trị kia thấp. Tuy nhiên một số người dùng có lựa chọn khác, đặt giá trị cao tại một nơi, thấp tại nơi khác. Trong khai phá dữ liệu, điểm rải không có ích, khi có quá nhiều điểm. Các điểm tạo thành đường có thể có ích trong một số trường hợp, như đánh giá mức đậm đặc hai chiều.

Một dạng hiển thị chuẩn khác dùng cho hai biến đồng thời là hiện các giá trị của các biến khác nhau theo tiến triển của thời gian. Đây là cách hiệu quả để phát hiện xu thế của hành vi, chẳng hạn đường phát triển dân số theo đường bậc hai, hay đường lũy thừa.

2.4.5. Các công cụ hiển thị trên quá hai biến

Hiển thị *phẳng*, chẳng hạn trên mặt giấy, chỉ thuận tiện đối với dữ liệu hai chiều, không thuận tiện cho dữ liệu nhiều chiều. Người ta có thể lấy ra các cặp biến, để hiển thị theo mối quan hệ hai chiều của chỉ hai biến. Như vậy có nhiều ma trận điểm.

Đồ họa tương tác được dùng để hiển thị nhiều chiều. Người ta có thể xoay hướng nhìn để phát hiện cấu trúc, mối quan hệ. Có hệ thống cho phép xoay hình một cách ngẫu nhiên, cho phép người dùng lựa chọn cấu trúc quan tâm. Tuy nhiên điều này chưa đạt được nhiều kết quả; nên phương pháp dùng các điểm vẫn được sử dụng.

Điểm rải *mắt cáo*² cũng được dùng cho nhiều cặp hai biến. Thay vì hiển thị các điểm rải cho mỗi cặp hai biến, người ta cố định các cặp cụ thể, sinh ra một loạt các điểm rải, cho phép đối sánh với các cặp khác.

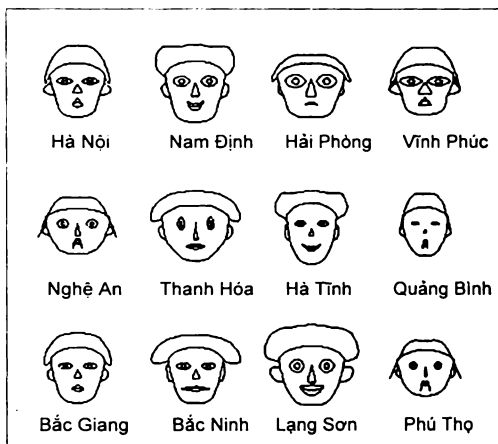
Một cách khác hẳn sử dụng các *biểu tượng*. Có sơ đồ nhỏ với nét khác nhau, tùy thuộc vào giá trị của biến. Biểu tượng ngôi sao

¹ Scatterplot: điểm rải.

² Trellis: mắt cáo.

là loại thông dụng. Các hướng khác nhau so với hướng gốc cho biết mức độ cao của biến.

Một dạng khác của điểm biểu tượng là *mặt Chernoff*. Tại các điểm có kích thước mặt hài hước, thể hiện qua độ dài của mũi, mức độ cười của môi, dạng của mắt... cho phép thể hiện giá trị biến. Trong mấy năm nay, nhận dạng sinh trắc, qua mặt, mắt người dùng được quan tâm.



Hình 2.9. Thể hiện qua mặt Chernoff

Các điểm theo trục song song cho thấy các biến theo trục song song, thể hiện mỗi trường hợp qua các đường điểm thẳng, nối đến giá trị đo của biến. Có thể sử dụng màu sắc để hiển thị giá trị biến. Người ta dùng thêm cả nét đường, điểm trục song song để hỗ trợ.

2.4.6. Phân tích thành phần chính

Các điểm rải sẽ chiếu dữ liệu nhiều biến lên không gian hai chiều, xác định theo hai biến. Việc này cho phép xét mối quan hệ hai biến của các biến, nhưng làm mất các mối quan hệ phức tạp. Người ta có thể chiếu theo nhiều hướng khác nhau, bằng cách dùng tổ hợp trọng số tuyến tính của các biến, chẳng hạn xác định hướng theo $2x_1 + 3x_2 + 4x_3$.

Với chỉ một ít biến, tìm không gian vừa ý theo cách thủ công là khả thi, bằng cách xoay phân phối dữ liệu. Khi nhiều biến, cách tốt nhất là để máy tính tự động tìm; người ta cần xác định hướng quan tâm rồi máy tính làm tiếp. Các phương pháp *theo hướng*¹ tuân theo nguyên tắc này. *Hướng* quan tâm trong các bài toán được xem là thuận lợi cho việc phát hiện, tức hướng cho phép dữ liệu theo các hướng đã chọn này có tác động nhiều đến vấn đề quan tâm.

Bài toán đặt ra trong phần này là phát hiện những thay đổi thực chất trong dữ liệu. Đây là cách để giảm chiều của bộ dữ liệu, hoặc là để diễn giả dữ liệu, hoặc là tránh trùng lặp và chuẩn bị cho phân tích.

Giả sử X là ma trận dữ liệu $n \times p$, dòng thể hiện trường hợp, ứng với vector dữ liệu $x(i)$. Cũng có thể coi đó là x^T , do coi vector dữ liệu như vector cột $p \times 1$. Giả sử X tập trung về trung bình, giá trị mỗi biến có liên quan đến trung bình mẫu của biến, tức trung bình đánh giá sẽ được rút ra từ mỗi cột.

Lấy a là vector cột $p \times 1$ là trọng số chiếu, chưa biết, gây nên thay đổi lớn nhất khi dữ liệu X được chiếu theo a . Việc chiếu của vector dữ liệu x bất kì là tổ hợp tuyến tính $a^T x = \sum_{j=1}^p a_j x_j$. Có thể thể hiện các giá trị chiếu vào a đối với tất cả các vector dữ liệu trong X qua Xa . Xa có kích thước $n \times p$ nhân với $p \times 1$, được $n \times 1$, vector cột các giá trị chiếu. Tiếp theo, người ta có thể xác định phương sai theo $\sigma_a^2 = (Xa)^T (Xa) = a^T X^T X a = a^T V a$, trong đó $V = X^T X$, là ma trận hiệp phương sai $p \times p$ của dữ liệu, do X có trung bình không. Do vậy, người ta có thể diễn tả phương sai của dữ liệu chiếu cần cực đại hóa, σ_a^2 là hàm của cả a và ma trận hiệp phương sai V .

Việc cực đại hóa σ_a^2 không dễ xác định, do người ta có thể tăng σ_a^2 không giới bằng cách tăng kích thước thành phần a . Có thể đặt ra một số ràng buộc, vậy người ta đặt ràng buộc chuẩn trên

¹ Projection pursuit method: phương pháp chiếu, theo hướng.

các vector a để $a^T a = 1$. Với ràng buộc này, có thể viết lại bài toán tối ưu cực đại hóa $u = a^T V a - \lambda (a^T a - 1)$, trong đó λ là nhân tử Lagrange. Sai phân thu được $\frac{\partial u}{\partial a} = 2Va - 2\lambda a = 0$, đưa về dạng giá trị riêng quen thuộc $(V - \lambda I)a = 0$.

Do vậy thành phần chính đầu tiên a là vector riêng liên quan đến giá trị riêng của ma trận hiệp phương sai V . Hơn nữa, thành phần thứ hai, hướng trực giao với thành phần thứ nhất để có thay đổi hình chiếu lớn nhất, là vector riêng ứng với giá trị riêng lớn nhất của V , và tiếp, vector riêng đối với giá trị riêng lớn nhất thứ k ứng với hướng thành phần chính thứ k .

Có kĩ thuật thống kê tên là *phân tích nhân tố*¹ thường bị lẫn với phân tích thành phần chính, tuy hai kĩ thuật có mục đích khác hẳn. Phân tích thành phần chính nhằm chuyển dữ liệu về các biến mới. Người ta có thể chọn đúng vài biến vẫn đảm bảo mô tả dữ liệu. Phân tích nhân tố là một mô hình đối với dữ liệu, dựa trên kĩ pháp cho phép xác định các biến được đo $X_1, \dots, X_p$ như tổ hợp tuyến tính của một số m , $m < p$, nhân tố chưa rõ ràng, tức các biến không thể đo một cách hiển hiện. Đích của phân tích nhân tố là tìm ra thông tin về các biến chưa rõ này.

Người ta có thể xác định vector cột $m \times 1$, $F = (F_1, \dots, F_m)$ của các biến chưa rõ, nhận các giá trị $f = (f_1, \dots, f_m)$. Khi đó vector dữ liệu được đo $x = (x_1, \dots, x_p)^T$, vector cột $p \times 1$, được coi như hàm tuyến tính của f : $x = \Lambda f + e$; trong đó Λ là ma trận $p \times m$ là tải nhân tố², mang các trọng số cho nhân tố đối với mỗi biến thể hiện. Các thành phần của vector $p \times 1$, e là các biến ngẫu nhiên không có tương quan; đôi khi được gọi là nhân tố đặc biệt, vì chúng chỉ tác động với biến thể hiện, tức biến quan sát X_j , $j \in [1, p]$. Phân tích nhân tố là trường hợp đặc biệt của các mô hình quan hệ tuyến tính có cấu trúc, sẽ trình bày sau; nên không đi sâu ở đây. Tuy nhiên do phân tích nhân tố là cấu trúc mô hình trước tiên về dạng được phát triển

¹ Factor analysis: phân tích nhân tố.

² Factor loading: tải nhân tố.

này, nó có vị trí đặc biệt không chỉ do tính lịch sử, mà nó được sử dụng nhiều trong các mô hình.

Phân tích nhân tố được đề cập nhiều, một phần do các giải pháp của nó hay thay đổi sang dạng khác. Công thức nhân tố mới $x = Af + e$ nêu trên được xác định qua các ma trận trực giao M , sao cho $x = (AM)(Mf) + e$. Điều này ứng với việc xoay nhân tố trong không gian. Do vậy, các nhân tố rút ra không duy nhất, trừ phi sử dụng các ràng buộc. Có nhiều ràng buộc, có phương pháp cho phép rút ra nhân tố để trọng số sát với 0 hay 1, hay xác định các biến rõ như xác định tập con của các nhân tố.

2.4.7. Thang đo nhiều chiều

2.4.7.1. Giới thiệu về thang đo nhiều chiều

Phần trên đã trình bày phân tích thành phần chính để chiếu dữ liệu nhiều biến sang mặt phẳng, mà vẫn đảm bảo dữ liệu ít bị biến dạng. Điều này cho phép người ta kiểm tra dữ liệu trực quan, tuy mất một chút thông tin. Phương pháp này chỉ hiệu quả khi dữ liệu thuộc không gian con tuyến tính hai chiều, của không gian của các biến được đo. Tuy dữ liệu có vẻ phẳng, nhưng hoàn toàn khác trong không gian nhiều chiều. Do vậy phân tích thành phần chính sẽ sai trong không gian hai chiều. Các phương pháp theo thang đo nhiều chiều tìm cách thể hiện các điểm dữ liệu trong không gian có số chiều ít hơn, vẫn đảm bảo khoảng cách giữa các điểm dữ liệu.

2.4.7.2. Thang đo nhiều chiều

Có nhiều phương pháp thang đo nhiều chiều, tùy theo (i) cách xác định khoảng cách; (ii) khớp khoảng cách với dữ liệu; và (iii) cách thực hiện tính toán. Phân tích thành phần chính có thể không quan tâm đến dạng khoảng cách cơ bản. Theo tiếp cận này, các khoảng cách giữa các điểm dữ liệu theo Oclit, hay Pitago, và chúng khớp với khoảng cách trong không gian thu hẹp có sử dụng độ đo Oclit. Tổng bình phương khoảng cách giữa các điểm dữ liệu ban đầu đến các hình chiếu của chúng đảm bảo số đo về chất

lượng thể hiện. Các phương pháp thang đo khác cũng liên quan đến độ đo về chất lượng thể hiện.

Do các phương pháp thang đo nhiều chiều vẫn cần bảo toàn khoảng cách, nên điều quan tâm đầu tiên trong phân tích dữ liệu là bảo toàn khoảng cách. Do vậy, chưa cần biết giá trị được đo của các biến đối tượng thế nào, mà chỉ xem các đối tượng giống nhau ra sao. Phương pháp thang đo nhiều chiều được dùng trong nhiều lĩnh vực, như đo về tâm lý, nghiên cứu thị trường, tìm hiểu về nhận thức mối quan hệ giữa các đối tượng.

2.4.7.3. Phương pháp tọa độ chính

Từ ma trận X $n \times p$, người ta có thể tính được ma trận $n \times n$ $B = XX^T$. Do các thang đo này là $O(n^2)$ theo cả bộ nhớ và thời gian, tiếp cận này không thực tế khi số các đối tượng n quá lớn. Còn theo khoảng cách Oclit giữa đối tượng i và j là $d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$

Nếu người ta chuyển ngược mối quan hệ này thì khi có các khoảng cách D , suy ra từ các điểm dữ liệu ban đầu theo khoảng cách Oclit hay theo cách khác, người ta có thể tính được các phần tử của B . Phân rã B để được các tọa độ của các điểm. Một cách phân rã B là theo vectơ riêng. Nếu chọn những cái liên quan đến hai giá trị riêng lớn nhất, người ta sẽ có thể hiện hai chiều, đảm bảo giữ được cấu trúc của dữ liệu. Tính khả thi của thủ tục này tùy thuộc vào khả năng chuyển ngược công thức vừa nêu, điều này thực hiện được nếu có thêm điều kiện. Do dịch chuyển trung bình và xoay cấu hình của các điểm không ảnh hưởng đến khoảng cách giữa các điểm, đối với bất kì tập các khoảng cách sẽ có vô hạn giải pháp, khác với cấu hình lại vị trí và hướng các điểm.

Ràng buộc đủ để thực hiện việc trên là giả thiết về trung bình các biến là 0. Người ta giả thiết $\sum_i x_{ik}, k = 1..p$. Điều này có nghĩa

$\sum_i b_{ij} = \sum_j b_{ij} = 0$. Giả sử trong công thức cân đảo, trước tiên theo i , sau đó theo j , người ta thu được $\sum_i d_{ij}^2 = tr(B) + nb_{jj}$;

$\sum_j d_{ij}^2 = \text{tr}(B) + nb_{ii}$; $\sum_{i,j} d_{ij}^2 = 2\text{tr}(B)$; trong đó $\text{tr}(B)$ là vết của ma trận B . Đẳng thức cuối thể hiện $\text{tr}(B)$ theo thuật ngữ của d_{ij} ²; đẳng thức thứ nhất, thứ hai diễn tả b_{ii} , b_{ii} theo thuật ngữ d_{ij} ² và $\text{tr}(B)$, và theo riêng d_{ij} ². Đưa thông tin này đến công thức cân đảo trên, người ta thu được kết quả.

Quá trình này sử dụng phương pháp *tọa độ chính*¹. Nó có thể cho thấy đích theo tính toán thành phần từ phân tích thành phần chính của ma trận X , tức một phần tử của ma trận XX^T , là như các tọa độ của phân tích thang đo vừa nêu trên. Tuy nhiên, nếu ma trận B không hiện lên dạng XX^T , mà theo dạng khác như hiệu của giữa các cặp đối tượng, không đảm bảo được rằng các giá trị riêng là không âm. Có thể bỏ qua các giá trị riêng âm nhỏ.

Thang đo nhiều chiều truyền thống theo hai chiều tìm thấy chiếu vào không gian hai chiều chính xác nhất theo nghĩa $\sum \sum (\delta_{ij} - d_{ij})^2$, theo i, j ; trong đó δ_{ij} là khoảng cách quan sát thấy giữa các điểm i và j trong không gian p chiều, và d_{ij} trong không gian hai chiều. Khi cho khoảng cách và độ khác nhau có được theo cách này hay cách khác, người ta có thể tìm phân phối các điểm trong không gian hai chiều, mà cực tiểu tổng độ lệch bình phương $\sum \sum (\delta_{ij} - d_{ij})^2$. Người ta nói lòng hạn chế về cân có cấu hình sau khi chiếu. Với điều kiện này, không có nghiệm đại số chính xác, nên cần sử dụng phương pháp số: dùng hàm $2n$ tham số, tức các tọa độ của các điểm trong không gian hai chiều, cần được tối thiểu hóa. Hàm đích $\sum \sum (\delta_{ij} - d_{ij})^2$ ứng với khoảng cách giữa các điểm tốt ra sao theo cấu hình khớp với đầu bài, là bất biến với phép xoay và chuyển dịch. Tất nhiên nó không bất biến với việc chia lại thang đo: nếu δ_{ij} được bội lên với hằng số, người ta có thể có cùng lời giải, nhưng kết quả khác với $\sum \sum (\delta_{ij} - d_{ij})^2$. Để cho phép các tình huống khác được so sánh, người ta chia $\sum \sum (\delta_{ij} - d_{ij})^2$ cho $\sum \sum d_{ij}^2$, thu được tổng bình phương chuẩn hóa. Hàm đích thông dụng là căn bậc hai của đánh giá đó, gọi là *nhấn*². Một dạng khác của sự

¹ Principal coordinate: tọa độ chính.

² Stress: căng, nhấn.

nhân là $\sqrt{\frac{\sum_i \sum_j (\delta_{ij}^2 - d_{ij}^2)^2}{\sum_i \sum_j d_{ij}^4}}$. Độ đo này giả sử sai khác giữa độ lệch

ban đầu và khoảng cách trong cấu hình hai chiều phụ thuộc vào sự không nhất quán ngẫu nhiên và độ vận tùy ý, tức $d_{ij} = \delta_{ij} - \varepsilon_{ij}$. Có thể xây dựng được các mô hình lí tưởng hơn.

2.4.7.4. Nhận xét về thang đo nhiều chiều

Các phương pháp thang đo nhiều chiều nhắm vào mô hình hóa sự khác nhau; được gọi là các phương pháp độ đo. Đôi khi cần đến tiếp cận tổng quát. Chẳng hạn không cho sự giống nhau chính xác, chỉ có thứ tự hạng mục, như đối tượng A và B là giống nhau hơn B và C...; hay chưa được chuẩn bị về mối quan hệ theo dạng cụ thể giữa δ_{ij} và d_{ij} mà chỉ có quan hệ đơn điệu. Điều này cần đến tiếp cận hai giai đoạn, nhưng với kĩ thuật như hồi quy đơn điệu thay cho hồi quy tuyến tính đơn giản, người ta thu được thang đo nhiều chiều không đo được. Thuật ngữ *không đo được* ở đây cho biết phương pháp tìm cách bảo toàn chỉ mỗi quan hệ chuẩn tắc.

Thang đo nhiều chiều là phương pháp mạnh để hiển thị dữ liệu, để thấy được cấu trúc. Tuy nhiên với các phương pháp đồ họa khác vừa mô tả, khi có nhiều điểm dữ liệu, cấu trúc trở nên rối rắm. Ngoài ra, do thang đo nhiều chiều đòi hỏi áp dụng phép chuyển hóa lí tưởng đối với dữ liệu, ngoài phép phân tích thành phần chính, thể hiện dữ liệu rải. Thực tế trong vài tình huống, sự khác nhau giữa các đối tượng có thể được xác định chính xác hơn khi các đối tượng là giống nhau so với khi các đối tượng khác nhau. Chẳng hạn về đối tượng được sản xuất. Các đối tượng được sinh ra trong thời gian gần thường có nhiều điểm chung, còn đối tượng xa nhau về thời gian sẽ có ít điểm chung. Có thể thể hiện điều này trên đồ thị điểm rải. Hiện tượng này được gọi là hiệu ứng *hình móng ngựa*¹.

¹ Horseshoe effect: hiệu ứng hình móng ngựa.

Các phương pháp thang đo nhiều chiều hiển thị điển hình các điểm dữ liệu trong không gian hai chiều. Nếu các biến cũng được mô tả trong không gian này, tức dữ liệu dạng vectơ, mối quan hệ giữa các điểm dữ liệu và các biến là dễ thấy. Cho mỗi quan hệ phi tuyến phức tạp không gian do các biến đã cho lúc ban đầu xác định và không gian dùng để hiển thị dữ liệu, việc thể hiện các biến ban đầu không dễ dàng. Các điểm hiển thị cả những điểm dữ liệu lẫn các biến được gọi là *điểm kép*¹. *Kép* được hiểu là có hai kiểu được hiển thị, chứ không phải hai chiều. Do vậy, có thể phát triển dạng điểm kép ba. Các dạng thang đo nhiều chiều cần đến sự chuyển hóa phi tuyến sẽ sinh ra điểm kép phi tuyến. Điểm kép đã từng được thể hiện qua miền trong các điểm. Việc diễn giải hiệu quả của hiển thị kép và nhiều chiều cần đến kinh nghiệm và thực tế.

2.4.8. Đánh giá hiệu năng

Khi xem xét việc đánh giá hiệu năng của phân tích dữ liệu, trước tiên người ta xét nhiệm vụ phân tích cần được đánh giá. Nhiệm vụ này có thể là (i) trả lời *câu hỏi được xử lý theo kiểu một lô*, tức người dùng đưa câu hỏi và nhận trả lời; (ii) thực hiện các phiên tương tác, tức người dùng xác định nhu cầu thông tin qua một loạt các bước tương tác với hệ thống. Ngoài ra, nhiệm vụ phân tích cũng gồm tổ hợp của hai chiến lược này.

Ngoài bản chất của yêu cầu câu hỏi, người ta cần xét thông số thiết lập khi đánh giá và dạng giao diện sử dụng. (i) Liên quan đến thông số thiết lập, người ta *đánh giá về thực nghiệm* trong phòng thí nghiệm có thể hoàn toàn khác với đánh giá tại hiện trường. (ii) Liên quan đến loại giao diện, trong khi hệ thống thư viện cũ vẫn thao tác theo lô, các hệ thống mới hơn cho phép dùng *đồ họa chất lượng cao*, thường là hướng người dùng tới giao diện phức tạp, thường thao tác tương tác.

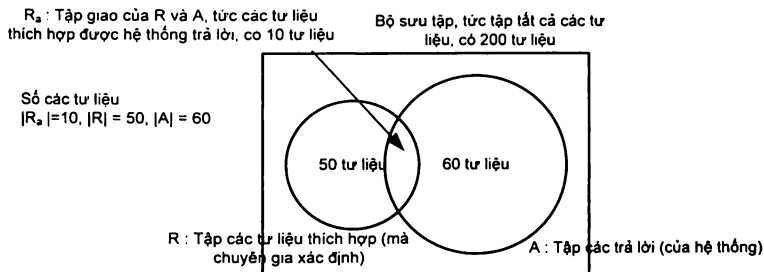
¹ Biplot: điểm kép.

Đánh giá hiệu năng phân tích dữ liệu trong thời gian đầu của các hệ thống phân tích dữ liệu trên máy tính thường tập trung vào các thực nghiệm trong phòng thí nghiệm, thiết kế cho các giao diện theo lô. Vào năm 1990, người ta quan tâm đến *đánh giá hiện trường*. Hai lí do chính khiến cần thiết đánh giá hiện trường là (i) khả năng lặp lại của đánh giá; và (ii) quy mô của phòng thí nghiệm là thông số đóng.

Phần này tập trung chủ yếu vào thực nghiệm trong các phòng thí nghiệm. Kiểu hỏi ở đây chỉ là theo lô; còn kiểu tương tác được trình bày trong phần sau.

2.4.8.1. Độ nhạy và chính xác

Xét yêu cầu thông tin thí dụ I , tức tập các tham chiếu thử nghiệm, và tập R các tư liệu thích hợp. Cho $|R|$ là số các tư liệu trong tập R này. Giả sử chiến lược tìm kiếm đã biết, đang được đánh giá, xử lí yêu cầu thông tin I , và tạo nên tập trả lời tư liệu A . Cho $|A|$ là số các tư liệu trong tập A . Ngoài ra $|R_a|$ là số các tư liệu trong giao của hay tập R và A .



Hình 2.10. Hình minh họa để tính độ nhạy và độ chính xác

- *Độ nhạy* là phần các tư liệu thích hợp, tức tập R , mà đã tìm thấy, tức được thể hiện trong R_a . Độ nhạy $R_c = |R_a| / |R|$;
- *Độ chính xác* là phần tư liệu đã tìm thấy, tức tập A , mà thích hợp, tức thể hiện trong R_a . Độ chính xác $P = |R_a| / |A|$.

Theo hình minh họa $R_c = 10 / 50 = 1 / 5$; $P = 10 / 60 = 1 / 6$.

Người ta cũng hiểu về độ nhạy và độ chính xác theo minh họa khác. Xét bảng dữ liệu hai chiều, trong đó các cột ứng với sự thích hợp, hàng ứng với tư liệu tìm thấy hay không tìm thấy.

Bảng 2.6. Xác định các giá trị để đánh giá độ nhạy và chính xác

	Thích hợp	Không thích hợp	← Theo chuyên gia
Tìm thấy (+)	đúng thấy (tp)	sai thấy (fp)	
Không tìm thấy (-)	sai không thấy (fn)	đúng không thấy (tn)	

↗ Trả lời của hệ thống
 ↖ Ý kiến thứ nhất: đánh giá về chiến lược của hệ thống

Theo xác định trên:

1. tp là các tư liệu tìm thấy và thích hợp, tức R_A ; $tp = 10$;
2. fp là các tư liệu tìm thấy, nhưng chuyên gia không cho là thích hợp; đây là hiện tượng sai của chiến lược tìm kiếm; $fp = 60 - 10 = 50$;
3. fn ứng với hiện tượng sai của chiến lược tìm kiếm, tức không thấy những tư liệu được chuyên gia coi là thích hợp; $fn = 50 - 10 = 40$;
4. tn ứng với sự kiện đúng, do hệ thống không trả lời những tư liệu mà chuyên gia không cho là thích hợp; tn là phần bù của tập R và A , tức $tn = 200 - (50 + 60 - 10) = 100$.

Người ta xác định:

- $R_c = tp / (tp + fn) = 10 / (10 + 40) = 1 / 5$;
- $P = tp / (tp + fp) = 10 / (10 + 50) = 1 / 6$.

Người ta còn đưa ra khái niệm *chính xác kết quả*¹, cho biết kết quả phân lớp đúng, được xác định như sau: $Ac = (tp + tn) / (tp + fp + fn + tn) = (10 + 100) / (10 + 50 + 40 + 100)$, tức có 110 tư liệu được chiến lược tìm kiếm đúng, hay phân lớp, trong số 200 tư liệu tổng cộng.

¹ *Accuracy* là chính xác, của kết quả, thí dụ bắn trúng vòng 10. *Precision* là chính xác về độ đo, thí dụ hai lần bắn đều sát nhau.

Khái niệm về độ nhạy và độ chính xác giả sử tất cả các tư liệu trong tập trả lời A đã được kiểm tra. Tuy nhiên, người dùng thường không trình bày tất cả tư liệu trong tập A ngay một lúc. Thay vì thể hiện hết các tư liệu, trước tiên tập A được xếp theo độ thích hợp. Rồi người dùng xét danh sách bắt đầu từ tư liệu đầu tiên. Trong tình huống này, độ nhạy và độ chính xác thay đổi theo việc rà soát tập trả lời A . Do đó, đánh giá đúng yêu cầu vạch đường cong của độ chính xác với độ nhạy.

Xét sự tập tham chiếu và những tập yêu cầu thông tin thí dụ của nó. Hãy tập trung vào yêu cầu thông tin thí dụ đã cho, tạo nên câu hỏi q . Giả sử tập R_q chứa các tư liệu thích hợp đối với q đã được xác định. Không mất tính tổng quát, giả sử tập R_q gồm các tư liệu sau $R_q = \{d3, d5, d9, d25, d39, d44, d56, d71, d89, d123\}$. Vậy, theo nhóm các chuyên gia, có 10 tư liệu thích hợp với câu hỏi q .

Xét thuật toán tìm kiếm mới; giả sử nó trả về, đối với câu hỏi q , xếp hạng của các tư liệu trong tập trả lời như sau:

Bảng 2.7. Xếp hạng tập trả lời

d123	d9	d38
d84	d511	d48
d56	d129	d250
d6	d187	d113
d8	d25	d3

Bảng 2.8. Tính độ phủ và độ chính xác

Tư liệu đã xếp	Độ phủ, nhạy ($ R_s / R $), $ R = 10$	Độ chính xác ($ R_s / A $), $ A $ xuất hiện
d123	1/ 10	1/ 1
d84	1/ 10	1/ 2
d56	2/ 10	2/ 3
d6	2/ 10	2/ 4
d8	2/ 10	2/ 5

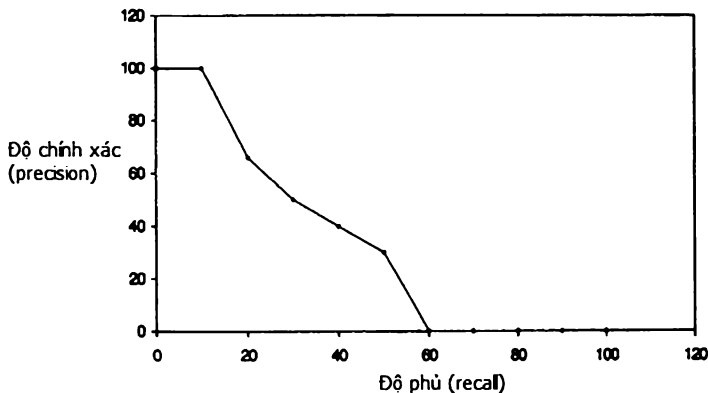
Tư liệu đã xếp	Độ phủ, nhảy ($ R_s / R $), $ R = 10$	Độ chính xác ($ R_s / A $), $ A $ xuất hiện
d9	3/ 10	3/ 6
d511	3/ 10	3/ 7
d129	3/ 10	3/ 8
d187	3/ 10	3/ 9
d25	4/ 10	4/ 10
d38	4/ 10	4/ 11
d48	4/ 10	4/ 12
d250	4/ 10	4/ 13
d113	4/ 10	4/ 14
d3	5/ 10	5/ 15

Các tư liệu thích hợp với câu hỏi q được tô đậm. Nếu xét xếp hạng này, từ tư liệu đầu tiên, người ta có quan sát sau:

- Tư liệu *d123* được xếp hạng đầu. Ngoài ra tư liệu này ứng với 10% tất cả các tư liệu thích hợp trong R_q . Vậy, người ta nói rằng có độ chính xác 100% ở 10% độ phủ, hay độ phủ;
- Ứng với dòng thứ ba, *d56*, người ta có độ chính xác 66% ở 20% độ phủ;
- Nếu tiến hành xem xét xếp hạng được tạo, người ta có thể vạch đường cong như trong hình. Độ chính xác ở mức độ phủ trên 50% xuống dần tới 0 vì không phải tất cả những tư liệu thích hợp được tìm thấy. Chính xác này thường dựa trên mức *phủ chuẩn* 11, thay vì 10. Mức phủ chuẩn là 0%, 10%, ..., 100%. Đối với mức phủ 0%, độ chính xác thu được nhờ nội suy.

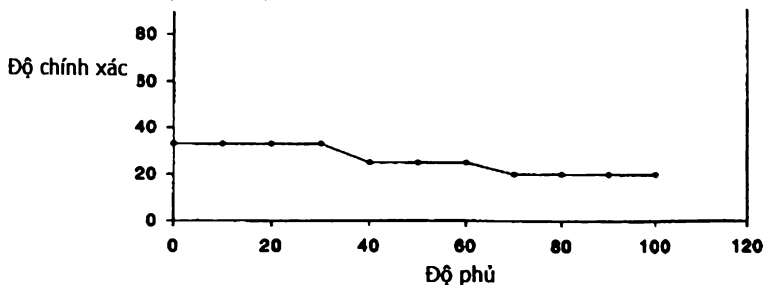
Trong thí dụ trên, độ chính xác và độ phủ là cho câu hỏi đơn. Thông thường, tất nhiên thuật toán tìm kiếm được đánh giá nhờ việc khai thác vài câu hỏi khác nhau. Trong trường hợp này, đối với mỗi câu hỏi, đường cong chính xác với độ phủ được tạo. Để đánh giá hiệu năng tìm kiếm của thuật toán theo tất cả các câu hỏi thử nghiệm, người ta lấy trung bình độ chính xác đối với mỗi mức

độ phủ $\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$; trong đó $\bar{P}(r)$ là chính xác trung bình tại mức độ nhảy r ; N_q là số các câu hỏi sử dụng; $P_i(r)$ là độ chính xác của câu hỏi i tại mức độ phủ r .



Hình 2.11. Độ chính xác tại độ phủ chuẩn 11

Do mức độ phủ đối với mỗi câu hỏi có thể khác nhau 11 mức phủ chuẩn, việc sử dụng thủ tục nội suy thường là cần thiết. Lại xét tập 15 tư liệu được xếp hạng trên. Giả sử tập các tư liệu thích hợp q thay đổi, nay là $R_q = \{d3, d56, d129\}$. Lúc này tư liệu thích hợp đầu tiên đối với q là $d56$, ứng với độ phủ 33%, với độ chính xác 33%, do một phần ba tổng các tư liệu thích hợp đã thấy. Tư liệu thích hợp thứ hai là $d129$, cho độ phủ 66.6% (tức 2/3), với độ chính xác 25% (tức 2/8).

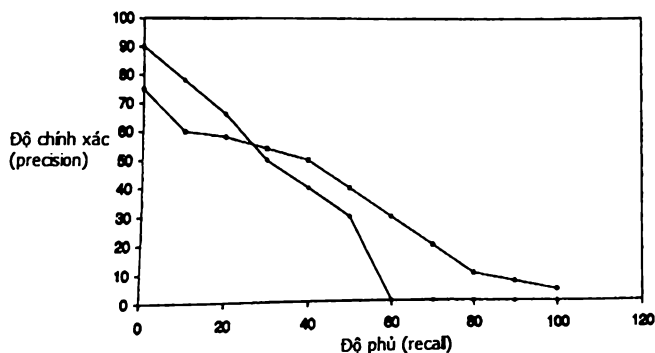


Hình 2.12. Độ chính xác nội suy tại các mức phủ chuẩn 11, liên quan đến $R_q = \{d3, d56, d129\}$.

Việc nội suy thực hiện: cho $r_j, j = 0, 1, 2, \dots, 10$ là tham chiếu đến mức phủ chuẩn thứ j , tức r_s tham chiếu đến mức phủ 50%. Vậy $P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$. Khẳng định độ chính xác nội suy tại mức phủ chuẩn thứ j là cực đại của độ chính xác đã biết tại bất kì mức phủ nào giữa mức thứ j và $j+1$.

Trong thí dụ cuối cùng, luật nội suy đạt được độ chính xác và độ phủ như trong hình. Tại các mức độ phủ 0%, 10%, 20% và 30%, độ chính xác là 33%, là độ chính xác tại ngưỡng phủ 33.3%. tại mức phủ 40%, 50% và 60%, độ chính xác nội suy là 25%, chính là chính xác tại mức phủ 66.6%. Tại mức phủ 70%, 80%, 90% và 100%, độ chính xác nội suy là 20, chính là độ chính xác tại độ phủ 100%.

Đường cong về độ chính xác với độ phủ dựa trên việc lấy trung bình kết quả của các câu hỏi khác nhau thường được tham chiếu như độ chính xác đối với độ phủ trong hình. Các đường trung bình được dùng bình thường để so sánh hiệu năng tìm kiếm, phân loại của các thuật toán khác nhau. Hiện tại, người ta có thể so sánh hiệu năng tìm kiếm của thuật toán mới để xuất so với hiệu năng tìm kiếm mô hình không gian véc tơ cổ điển. Hình sau cho thấy độ chính xác trung bình đối với dạng độ phủ của hai thuật toán khác nhau. Hình cho thấy thuật toán thứ nhất có độ chính xác cao hơn tại độ phủ thấp; thuật toán thứ hai có độ chính xác cao hơn tại độ phủ cao.



Hình 2.13. Độ phủ trung bình đối với dạng độ chính xác đối với hai thuật toán

Một tiếp cận khác cho phép tính độ chính xác trung bình tại các giá trị ngắt tự liệu¹. Người ta có thể tính độ chính xác trung bình khi 5, 10, 15, 20, 30, 30 hay 100 tự liệu thích hợp tìm thấy. Thủ tục tương tự như tính độ chính xác tại mức phù chuẩn 11, nhưng cung cấp thông tin phụ về hiệu năng tìm kiếm của thuật toán xếp hạng.

Độ chính xác trung bình đối với đáng độ phù là chiến lược đánh giá chuẩn cho các hệ thống tìm kiếm thông tin và được sử dụng mở rộng trong các tài liệu về tìm kiếm. Chúng có ích do cho phép người ta đánh giá lượng hóa cả chất lượng của tập trả lời toàn thể, lẫn sự trải rộng của thuật toán tìm kiếm. Ngoài ra chúng đơn giản về trực quan, và có thể kết hợp trong đường cong đơn. Tuy nhiên, độ chính xác đối với đáng độ phù cũng có nhược điểm và bị phê phán vì đã ứng dụng rộng rãi. Tiếp sau đây là các kĩ thuật tóm lược độ chính xác đối với các đáng độ phù bằng giá trị số đơn.

Tóm lược giá trị đơn: độ chính xác trung bình đối với các đáng độ phù có ích để so sánh hiệu năng tìm kiếm của các thuật toán khác nhau trên một tập câu hỏi thí dụ. Tuy nhiên, có trường hợp mà người ta muốn so sánh hiệu năng tìm kiếm trên các câu hỏi riêng. Lí do là (i) độ chính xác trung bình trên nhiều câu hỏi có thể che mất lỗi quan trọng trong thuật toán nghiên cứu; (ii) khi so sánh hai thuật toán, người ta muốn biết thuật toán này có hơn hẳn thuật toán kia đối với mỗi câu hỏi của tập câu hỏi thí dụ. Lưu ý rằng sự kiện này có thể dễ bị che khuất khi tính độ chính xác trung bình. Trong các trường hợp này, người ta dùng giá trị chính xác đơn, đối với mỗi câu hỏi. Giá trị đơn này được diễn giải như tóm lược của độ chính xác tương ứng đối với đường cong độ phù. Thông thường, việc tóm lược giá trị đơn này được lấy như độ chính xác tại mức phù xác định. Ở đây, người ta đánh giá độ chính xác khi quan sát tự liệu thích hợp đầu tiên và lấy độ chính xác này như tóm lược giá trị đơn. Tất nhiên, đây không là tiếp cận tốt, và còn nhiều chiến lược lí thú khác.

¹ Document cutoff value: giá trị ngắt tự liệu.

Độ chính xác trung bình tại các tư liệu thích hợp đã thấy: ý tưởng ở đây là tạo tóm lược giá trị đơn về xếp hạng đáng điều về sự chính xác thu được sau khi quan sát thấy tư liệu thích hợp mới, trong xếp hạng. Ở đây, xét thí dụ cũ, với $R_q = \{d3, d5, d25, d39, d44, d56, d71, d89, d123\}$. Đáng điều của độ chính xác sau mỗi tư liệu thích hợp mới quan sát được là 1; 0.66; 0.5; 0.4; và 0.3. Vậy, độ chính xác trung bình tại những tư liệu thích hợp thấy được là $(1 + 0,66 + 0,5 + 0,4 + 0,3)/5 = 0,57$. Số đo này lợi cho hệ thống tìm kiếm nhanh các tư liệu thích hợp, tức sớm trong xếp hạng. Tuy nhiên, thuật toán có thể cho độ chính xác trung bình tốt tại các tư liệu thích hợp thấy được, nhưng chịu hiệu năng thấp theo thuật ngữ của độ phù chung.

Độ chính xác R

$R = A/B$.

1.5

1.0

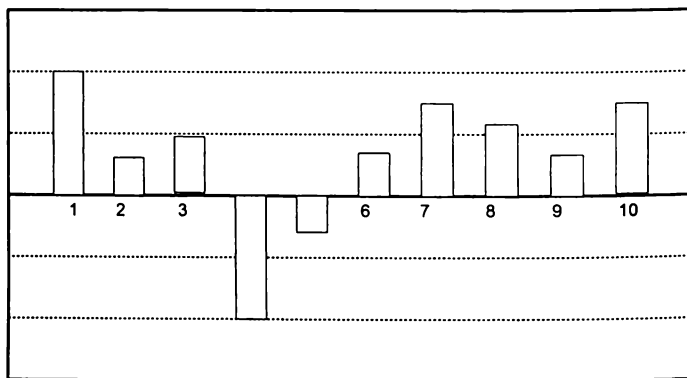
0.5

0.0

-0.5

-1.0

-1.5



Số các câu hỏi thống kê

Hình 2.14. Biểu đồ độ chính xác đối với 10 câu hỏi giả thuyết

Độ chính xác R: ở đây ý tưởng là tạo tóm lược giá trị đơn về xếp hạng bằng cách tính độ chính xác tại vị trí thứ R trong xếp hạng, trong đó R là số các tư liệu thích hợp đối với câu hỏi hiện tại, tức số các tư liệu trong tập R_q . Ở đây, xét tiếp thí dụ cũ. Giá trị của độ chính xác R là 0.4 đối với thí dụ đầu, do $R = 10$ và có 4 tư liệu thích hợp trong 10 tư liệu xếp hạng đầu tiên, và 0.33 đối với

thí dụ sau, do $R = 3$ và có 1 tư liệu thích hợp trong số 3 tư liệu xếp hạng đầu tiên. Số đo độ chính xác R là tham số có ích để quan sát hành vi của thuật toán đối với mỗi câu hỏi riêng trong thực nghiệm. Ngoài ra, người ta có thể tính đáng độ chính xác R trên tất cả các câu hỏi. Tất nhiên, việc dùng giá trị đơn để tóm lược toàn thể hành vi của thuật toán tìm kiếm trên vài câu hỏi có thể không hoàn toàn chính xác.

Biểu đồ độ chính xác: các số đo độ chính xác R đối với vài câu hỏi có thể dùng để so sánh lịch sử tìm kiếm của hai thuật toán như sau. Cho $RP_A(i)$ và $RP_B(i)$ là các giá trị độ chính xác R của thuật toán A, B đối với câu hỏi thứ i . Xác định $RP_{A/B}(i) = RP_A(i) - RP_B(i)$. Giá trị dương $RP_{A/B}(i)$ cho biết thuật toán của hiệu năng tốt hơn, đối với câu hỏi thứ i , và giá trị âm chỉ ra hiệu năng thuật toán B tốt hơn. Biểu đồ trong hình cho biết các giá trị $RP_{A/B}(i)$, với nhãn là độ chính xác $R_{A/B}$, đối với hai thuật toán tìm kiếm giả thuyết trên 10 câu hỏi thí dụ. Kiểu đồ thị này gọi là *biểu đồ độ chính xác* và cho phép người ta nhanh chóng so sánh lịch sử hiệu năng tìm kiếm của hai thuật toán qua đồ họa.

Thống kê bảng tóm lược: những số đo giá trị đơn có thể lưu trong bảng, cung cấp tóm lược về thống kê liên quan đến tập tất cả các câu hỏi trong nhiệm vụ tìm kiếm. Ở đây, các thống kê bảng tóm lược có thể gồm: số các câu hỏi sử dụng trong nhiệm vụ, tổng số các tư liệu mà các câu hỏi tìm thấy, tổng số tư liệu thích hợp được các câu hỏi tìm thấy...

Tính phù hợp của độ chính xác và độ phủ: độ chính xác và độ phủ được sử dụng để đánh giá hiệu năng tìm kiếm của thuật toán tìm kiếm. Tuy nhiên, xét cẩn thận hơn, người ta thấy có vấn đề với hai độ đo này.

1. *Đánh giá riêng của độ phủ* cực đại đối với các câu hỏi cần đến tri thức chi tiết về tất cả các tư liệu trong sưu tập. Với sưu tập lớn, tri thức này không sẵn sàng, nên không thể tính chính xác được độ phủ;

2. Độ chính xác và độ phù là *số đo liên quan* nhau, nắm các khía cạnh khác nhau của tập các tư liệu tìm thấy. Trong nhiều trường hợp, việc dùng độ đo đơn, kết hợp cả hai độ đo thì hợp lí hơn;
3. Độ chính xác và độ phù đo tính hiệu quả trên tập các câu hỏi được xử lí theo lô. Tuy nhiên, trong hệ thống hiện đại, tương tác là khía cạnh chính của quá trình tìm kiếm. Vậy nên, các độ đo định lượng có *tính hướng dẫn* quá trình tìm kiếm giờ đây phù hợp hơn;
4. Độ chính xác và độ phù dễ xác định khi bắt buộc tuân theo trật tự tuần tự đối với các tư liệu tìm kiếm được. Đối với hệ thống đòi hỏi xếp trật tự yếu hơn, độ chính xác và độ phù chưa đủ để đánh giá quá trình tìm kiếm.

2.4.8.2. Các độ đo khác

Do độ chính xác và độ phù, cho dù thông dụng, không luôn luôn là độ đo phù hợp nhất để đánh giá hiệu năng tìm kiếm, nên người ta đề xuất các độ đo khác.

- *Trung bình điều hòa*: người ta quan tâm đến độ đo kết hợp được cả độ chính xác lẫn độ phù. Một trong số đó là *độ đo trung bình điều hòa* F , được tính $F(j) = 2 / (\frac{1}{r(j)} + \frac{1}{P(j)})$, trong đó $r(j)$, $P(j)$ là độ phù, độ chính xác của tư liệu thứ j trong xếp hạng. Hàm F giả sử các giá trị trong $[0, 1]$. Nó là 0 khi không tìm thấy tư liệu thích hợp; là 1 khi tất cả tư liệu xếp hạng đều là thích hợp. Ngoài ra, trung bình điều hòa F lấy giá trị cao chỉ khi cả độ chính xác và độ phù là cao. Do đó quyết định giá trị cực đại của F được diễn giải như dự định tìm sự thỏa hiệp có thể nhất giữa độ chính xác và độ phù.
- *Độ đo E*: độ đo do Rijsbergen đưa ra là *độ đo đánh giá E*. Ý tưởng là cho phép người dùng xác định sự quan tâm hơn đối với độ chính xác hay đối với độ phù. E được xác định $E(j) = 1 - (1 + b^2) / (\frac{b^2}{r(j)} + \frac{1}{P(j)})$, trong đó $r(j)$, $P(j)$ là

độ phù, độ chính xác của tư liệu thứ j trong xếp hạng; b là tham số người dùng xác định, phản ánh mức quan trọng tương đối của độ chính xác và độ phù. Với $b = 1$, $E(j)$ như trung bình điều hòa $F(j)$. Với $b > 1$ cho biết người dùng quan tâm đến độ chính xác hơn độ nhạy; $b < 1$ ngược lại.

- *Độ đo hướng người dùng*: độ chính xác và độ phù dựa trên giả thiết rằng tập các tư liệu thích hợp đối với câu hỏi là như nhau, độc lập với người dùng. Tuy nhiên, người dùng khác nhau có lí giải khác nhau về tư liệu nào là thích hợp, tư liệu nào không. Để khắc phục vấn đề này, các *độ đo hướng người dùng* được dùng như tỉ lệ phù, tỉ lệ mới, độ phù tương đối và nỗ lực phù.

Trước hết, xét suu tập tham chiếu, yêu cầu thông tin I , chiến lược tìm kiếm cần đánh giá. Cho R là tập các tư liệu thích hợp và A là tập trả lời tìm thấy. Bên cạnh kí pháp này, có tập con của R , mà người dùng biết. Tập con là U , với lực lượng $|U|$. $R_k = U \cap A$ là các tư liệu người dùng biết, là các tư liệu thích hợp tìm thấy. $|R_k|$ là số các tư liệu trong R_k . Mô tả trực quan các tập này trong hình vẽ.



Hình 2.15. Các tỉ lệ mới và tỉ lệ phù đối với yêu cầu thông tin thí dụ

- *Tỉ lệ phù* được xác định như phần tư liệu mà người dùng đã biết là thích hợp mà đã tìm thấy: $\text{Độ phù} = |R_k| / |U|$;
- *Tỉ lệ mới* là phần các tư liệu thích hợp đã tìm thấy, mà người dùng chưa biết: $\text{Độ mới} = |R_u| / (|R_u| + |R_k|)$.

Tỉ lệ phủ cao cho biết hệ thống tìm thấy hầu hết các tư li mà người dùng mong đợi. Tỉ lệ mới và cao cho biết hệ thống hiện cho người dùng nhiều tư liệu thích hợp mới mà người dùng chưa biết trước.

Ngoài ra, hai độ đo khác cũng được xác định:

- *Độ phủ tương đối* là tỉ lệ giữa số các tư liệu thích hợp tìm thấy, bởi hệ thống, và số tư liệu thích hợp mà người dùng mong tìm thấy. Trong trường hợp mà người dùng tìm thấy nhiều tư liệu thích hợp như mong đợi, người ta dùng việc tìm và độ phủ tương đối là 1;
- *Nỗ lực phủ* là tỉ số giữa số các tư liệu thích hợp mà người dùng mong tìm thấy và số tư liệu được xét, với ý tìm kiếm tư liệu thích hợp mong muốn.

Các độ đo khác: các độ đo khác đáng quan tâm là *độ dài tìm kiếm mong muốn*, tốt để xử lí các tập tư liệu được xếp thứ tự yếu; *độ tin cậy*, chỉ tính đến các tư liệu thích hợp; *độ thất bại*, chỉ tính đến các tư liệu không thích hợp.

2.5. Phân phối dữ liệu

Trong lí thuyết xác suất và thống kê, phân phối xác suất là một hàm toán học. Nó cho biết xác suất xảy ra các sự kiện có thể khác nhau, trong thực nghiệm. Chẳng hạn biến ngẫu nhiên được sử dụng để biểu thị kết quả của việc tung đồng xu, tức thì thực nghiệm về đồng xu, $p(X = \text{mặt sấp}) = 0.5$ và $p(X = \text{mặt ngửa}) = 0.5$.

Phân phối xác suất là mô tả một hiện tượng ngẫu nhiên về xác suất các sự kiện. *Phân phối xác suất* được xác định dưới dạng không gian mẫu nền, là tập hợp của tất cả các đầu ra có thể xảy ra của hiện tượng ngẫu nhiên đang được quan sát. Không gian mẫu có thể là tập các số thực hoặc không gian vector lớn hơn, hoặc có thể là một danh sách các giá trị phi số.

Phân phối xác suất thường được chia thành hai lớp (i) *phân phối xác suất rời rạc*, áp dụng cho trường hợp mà tập hợp các d

ra có thể là rời rạc; chẳng hạn mã hoá bởi một danh sách rời rạc về xác suất đầu ra, gọi là *hàm mật độ xác suất*¹; (ii) *phân phối xác suất liên tục*, áp dụng cho các trường hợp mà tập hợp các đầu ra có thể có giá trị trong một dải liên tục, chẳng hạn như nhiệt độ vào một ngày nhất định thường được mô tả bởi hàm mật độ xác suất. Phân phối chuẩn thể hiện sự phân phối xác suất liên tục. Các thực nghiệm phức tạp hơn, chẳng hạn như các quá trình ngẫu nhiên xác định trong thời gian liên tục, có thể yêu cầu sử dụng các độ đo xác suất tổng quát hơn.

Một phân phối xác suất có không gian mẫu là tập các số thực được gọi là *đơn biến*, trong khi một phân phối có không gian mẫu là một không gian vector được gọi là *đa biến*². Một phân phối đơn biến cho xác suất của một biến ngẫu nhiên đơn, có các giá trị thay thế khác nhau. Một phân phối đa biến, phân phối xác suất kết hợp, cho biết xác suất của một vector ngẫu nhiên, tức danh sách của hai hoặc nhiều biến ngẫu nhiên, tham gia vào các tổ hợp của các giá trị khác nhau. Phân phối xác suất đơn biến quan trọng, thường gặp bao gồm (i) phân phối nhị thức; (ii) phân phối siêu hình học; (iii) phân phối chuẩn. Phân phối chuẩn nhiều biến thể là phân phối đa biến thường gặp.

2.5.1. Giới thiệu về phân phối dữ liệu

Để xác định phân phối xác suất cho các trường hợp đơn giản nhất, cần phân biệt giữa các biến ngẫu nhiên *rời rạc* và *liên tục*.

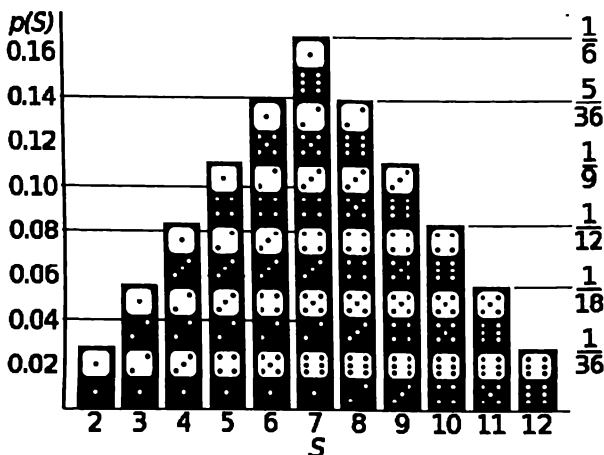
- Trong trường hợp rời rạc, chỉ cần xác định một hàm mật độ xác suất, gán cho xác suất của mỗi đầu ra có thể. Chẳng hạn tung con xúc xắc sáu mặt, xác suất mỗi mặt xuất hiện là $1/6$; xác suất các mặt chẵn là $p(2) + p(4) + p(6) = 1/6 + 1/6 + 1/6 = 1/2$;
- Trong trường hợp liên tục, khi biến ngẫu nhiên có các giá trị từ liên tục thì thông thường, bất kỳ đầu ra riêng lẻ nào

¹ Mass function: hàm khối, hàm mật độ.

² Multivariat: đa biến; univariat: đơn biến.

có xác suất bằng không và chỉ các sự kiện bao gồm vô số các đầu ra, chẳng hạn như các khoảng, có thể có xác suất dương. Chẳng hạn xác suất mà một vật nhất định cân nặng chính xác là 500 g là 0, do xác suất đo chính xác 500 g có xu hướng bằng không khi độ chính xác của dụng cụ đo tăng lên. Tuy nhiên, trong kiểm soát chất lượng, người ta có thể yêu cầu xác suất gói 500 g chứa từ 490 g đến 510 g không ít hơn 98%, và nhu cầu này ít nhạy hơn với độ chính xác của phép đo.

Xét các quân domino, trừ quân có hai mặt trắng, có tất cả 36 quân. Hàm mật độ xác suất PMF $p(S)$ xác định phân phối xác suất của tổng điểm S của hai mặt. Hình cho thấy $p(11) = 1/18$. Hàm PMF cho phép tính xác suất các sự kiện như $p(S > 9) = 1/12 + 1/18 + 1/36 = 1/6$, và tất cả các xác suất khác trong phân phối.



Hình 2.16. Hàm PMF với các quân domino

Phân phối xác suất liên tục có thể được mô tả bằng nhiều cách. Hàm mật độ xác suất mô tả xác suất của bất kỳ giá trị nào, và xác suất mà đầu ra nằm trong một khoảng nhất định có thể được tính bằng tích phân hàm mật độ xác suất trong khoảng thời gian đó. Mặt khác, hàm phân phối tích lũy mô tả xác suất mà biến ngẫu

nhiên không lớn hơn giá trị cho trước; xác suất mà kết quả nằm trong một khoảng nhất định có thể được tính bằng sai khác giữa các giá trị của hàm phân phối tích lũy tại các đầu mút của khoảng. Hàm phân phối tích lũy là hàm trái với hàm mật độ xác suất.

2.5.2. Một số thuật ngữ

Khi lí thuyết xác suất được sử dụng trong các ứng dụng khá đa dạng, thuật ngữ không thống nhất, hay gây nhầm lẫn. Các thuật ngữ sau được sử dụng cho các hàm phân phối xác suất không tích lũy:

- *Phân phối*¹, phân phối tần số: là một bảng hiển thị tần số các đầu ra khác nhau trong một mẫu;
- *Phân phối xác suất*²: là một bảng hiển thị xác suất của các đầu ra khác nhau trong một mẫu. Nó có thể là bảng phân phối tần số chuẩn hóa, trong đó tổng các xuất hiện của các đầu ra là 1;
- *Hàm phân phối*³: là một dạng hàm của bảng phân phối tần số;
- *Hàm phân phối xác suất*⁴: là một dạng hàm của bảng phân phối xác suất. Nó có thể là *hàm phân phối tần số chuẩn hóa*, ứng với miền dưới đồ thị bằng 1;
- *Mật độ xác suất*⁵, hàm mật độ xác suất, PMF, *hàm phân phối xác suất rời rạc*: thuật ngữ sử dụng cho các biến ngẫu nhiên rời rạc;
- *Phân phối phạm trù*⁶: dùng cho các biến ngẫu nhiên rời rạc với một tập các giá trị hữu hạn;

¹ Distribution: phân phối.

² Probabilty distribution: phân phối xác suất.

³ Distribution function: hàm phân phối.

⁴ Probability distribution function: hàm phân phối xác suất.

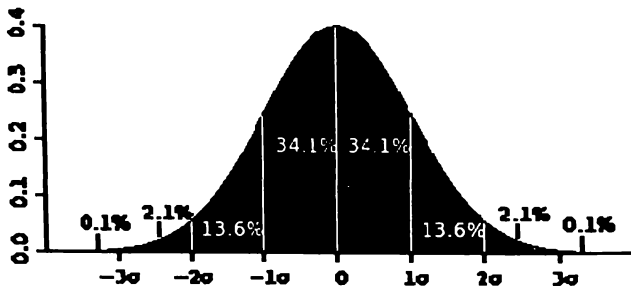
⁵ Probability mass: mật độ xác suất.

⁶ Categorical distribution: phân phối phạm trù.

- *Mật độ xác suất*¹, hàm mật độ xác suất, PDF, *hàm phân phối xác suất liên tục*: thuật ngữ dùng cho các biến ngẫu nhiên liên tục.

Có một vài thuật ngữ mơ hồ vì chúng tham chiếu đến phân phối không tích lũy hoặc tích lũy (i) *hàm phân phối xác suất*, dùng cho dữ liệu liên tục hay rời rạc, có tích lũy hay không tích lũy; (ii) *hàm xác suất*, có thể nhầm vào khái niệm đã nêu trên.

Trong hình là hàm mật độ xác suất, hay được gọi là đường hình chuông hay Gauss; đó là phân phối ngẫu nhiên liên tục quan trọng nhất. Các xác suất của các khoảng giá trị ứng với miền dưới đường cong.



Hình 2.17. Hàm mật độ xác suất PDF của phân phối chuẩn

Một số thuật ngữ cơ bản, liên quan đến dữ liệu trong phân tích thống kê và khai phá dữ liệu.

- *Mode*: đối với một biến ngẫu nhiên rời rạc, giá trị có xác suất cao nhất, tức vị trí mà hàm xác suất có đỉnh. Đối với một biến ngẫu nhiên liên tục, vị trí mà ở đó hàm mật độ xác suất có đỉnh;
- *Trợ giúp*: tập đóng nhỏ nhất mà phần bù của nó có xác suất 0;
- *Đầu*²: dài giá trị mà PMF hay PDF là tương đối cao;

¹ *Probability density*: mật độ xác suất.

² *Head*: đầu; *tail*: đuôi.

- *Đuôi*: phần bù của *đầu* trong *trợ giúp*. Đó là tập lớn các giá trị là PMF hay PDF tương đối nhỏ;
- *Giá trị kì vọng, hay trung bình*: trung bình có trọng số của các giá trị, sử dụng xác suất làm trọng số;
- *Trung vị*¹: giá trị sao cho tập các giá trị nhỏ hơn giá trị này có xác suất bằng một nửa;
- *Phương sai*²: là độ đo tán xạ của phân phối; nó là momen chuẩn hóa thứ hai của phân phối; trung bình là momen thứ nhất của phân phối;
- *Độ lệch chuẩn*³: căn bậc hai của phương sai; nó cũng là độ đo về tán xạ;
- *Đôi xiêng*: thuộc tính của một số phân phối, trong đó phần phân phối ở bên trái của một giá trị cụ thể là một hình ảnh phản chiếu của phần bên phải của nó;
- *Độ nghiêng*⁴: độ đo mức độ mà một PMF hay PDF có xu hướng lệch một bên so với trung bình của nó có ý nghĩa. Nó là momen chuẩn hóa thứ ba của phân phối;
- *Độ xiên*⁵: độ đo về tính đậm của đuôi của một PMF hay PDF. Đây là momen chuẩn hóa thứ tư của phân phối.

2.5.3. Hàm phân phối tích lũy

Bởi vì sự phân phối xác suất P trên trục thực được xác định bởi xác suất của một biến ngẫu nhiên vô hướng X trong $(-\infty, X]$, phân phối xác suất hoàn toàn được đặc trưng qua *hàm phân phối tích lũy*: $F(x) = P[X \leq x]$, $\forall x \in \mathbb{R}$.

¹ Median: điểm giữa, trung vị.

² Variance: phương sai.

³ Standard deviation: độ lệch chuẩn.

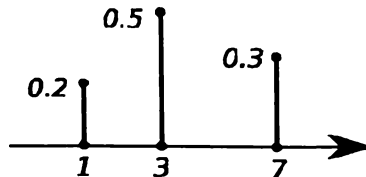
⁴ Skewness: độ nghiêng.

⁵ Kurtosis: độ xiên.

2.5.4. Phân phối xác suất rời rạc

Phân phối xác suất rời rạc là phân phối đặc trưng bởi hàm mật độ xác suất PMF. Do vậy, xác suất của biến ngẫu nhiên X là rời rạc.

X được gọi là biến ngẫu nhiên rời rạc, nếu $\sum_u P(X = u) = 1$, khi u nhận tất cả các giá trị có thể của X . Biến ngẫu nhiên rời rạc có thể giá sử chỉ hữu hạn hay vô hạn đến được các giá trị. Đối với số các giá trị tiềm năng được vô hạn đếm được, cho dù tổng xác suất là 1, các xác suất cần phải giảm xuống 0 đủ nhanh. Chẳng hạn nếu $P(X = x) = 1/2^n$, người ta có tổng xác suất là $1/2 + 1/4 + 1/8 + \dots = 1$.

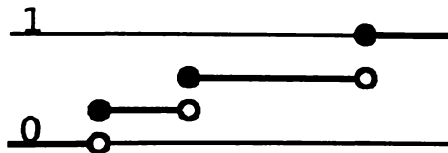


Hình 2.18. Hàm mật độ xác suất của phân phối xác suất rời rạc.

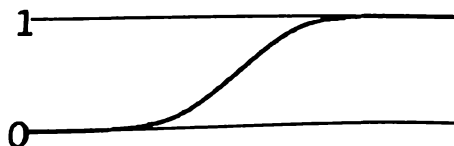
Xác suất của thành phần {1}, {3}, và {7} là 0.2, 0.5 và 0.3.

Tập không có điểm nào có xác suất 0

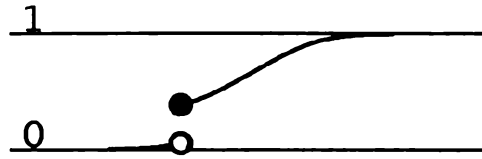
Các phân phối xác suất rời rạc nổi tiếng được sử dụng trong mô hình thống kê bao gồm (i) phân phối Poisson; (ii) phân phối Bernoulli; (iii) phân phối nhị thức; (iv) phân phối hình học; và (v) phân phối nhị thức âm. Ngoài ra, phân phối đều rời rạc thường được sử dụng trong các chương trình máy tính, cho phép tạo lựa chọn ngẫu nhiên với xác suất bằng nhau trong số các lựa chọn.



Hình 2.19. CDF của phân phối xác suất rời rạc



Hình 2.20. CDF của phân phối xác suất liên tục



Hình 2.21. CDF của phân phối xác suất cả liên tục lẫn rời rạc

2.5.4.1. Công thức lý thuyết về độ đo

Một hàm có thể đo được $X: A \rightarrow B$ giữa không gian xác suất $(A, \mathcal{A}_1, P)$ và không gian độ đo $(B, \mathcal{B}_1, P)$ được gọi là biến ngẫu nhiên rời rạc, cho thấy hình ảnh về tập đếm được và ảnh của tập thành phần được đo. Tức là $X^{-1}(b) \in \mathcal{A}_1$ đối với mọi $b \in B$. Nó yêu cầu suy ra hàm mật độ xác suất $f_X: X(A) \rightarrow \mathbb{R}$ thông qua $f_X(b) := P(X^{-1}(b))$. Vì các hình ảnh trước của các tập không giao nhau là không giao,

$$\sum_{b \in X(A)} f_X(b) = \sum_{b \in X(A)} P(X^{-1}(b)) = P\left(\bigcup_{b \in X(A)} X^{-1}(b)\right) = P(A) = 1.$$

Điều này khớp với định nghĩa đã đưa ra ở trên.

2.5.4.2. Mật độ tích lũy

Tương tự như trên, một biến ngẫu nhiên rời rạc có thể được định nghĩa như là một biến ngẫu nhiên có hàm phân phối tích lũy CDF chỉ tăng lên bằng cách nhảy các gián đoạn, tức là CDF của nó tăng chỉ khi nó nhảy lên giá trị cao hơn, và không đổi giữa các lần nhảy. Các điểm nhảy xảy ra chính xác tại các giá trị mà biến ngẫu nhiên có thể mất.

2.5.4.3. Thể hiện hàm delta

Phân phối xác suất rời rạc thường được thể hiện như hàm mật độ xác suất, dùng hàm delta Dirac, mà nhất quán với xử lý các phân phối rời rạc và liên tục. Việc này có ích khi xử lý phân phối xác suất có cả phần liên tục lẫn rời rạc.

2.5.4.4. Thể hiện hàm chỉ thị

Đối với biến ngẫu nhiên rời rạc X , lấy $u_0, u_1, \dots$ là các giá trị với giá trị xác suất khác 0. Kí hiệu $\Omega = X^{-1}(u_i) = \{\omega: X(\omega) = u_i\}$, $i = 0, 1, 2, \dots$

Đây là các tập không giao, nên $P(\bigcup_i \Omega_i) = \sum_i P(\Omega_i) = \sum_i P(X = u_i) = 1$. Do xác suất X nhận các giá trị khác với u_0 , u_1 là 0, nên người ta có thể viết X như $X(\omega) = \sum_i u_i 1_{\Omega_i}(\omega)$, ngoại trừ tập xác suất 0, trong đó 1_A là hàm chỉ thị của A . Đây có thể xem như định nghĩa khác về biến ngẫu nhiên rời rạc.

2.5.5. Phân phối xác suất liên tục

Phân phối xác suất liên tục là sự phân phối xác suất có một hàm phân phối tích lũy liên tục. Thông thường chúng được tạo ra từ hàm mật độ xác suất. Các nhà toán học gọi phân phối có hàm phân phối tích lũy liên tục là *hoàn toàn liên tục*, bởi vì hàm phân phối tích lũy của chúng là hoàn toàn liên tục đối với độ đo Lebesgue λ . Nếu hàm phân phối của X là liên tục, thì X được gọi là một *biến ngẫu nhiên liên tục*. Có nhiều phân phối xác suất liên tục (i) phân phối chuẩn; (ii) phân phối đều; (iii) phân phối χ^2 ...

Về thực quan, một biến ngẫu nhiên liên tục là một biến có thể lấy một dải giá trị liên tục; trái với một phân phối rời rạc, mà các giá trị có thể cho biến ngẫu nhiên nhiều nhất chỉ đến mức đếm được. Trong khi đối với phân phối rời rạc, không có sự kiện có xác suất bằng không, nhưng với phân phối liên tục, thì có thể. Chẳng hạn đo độ rộng của lá hoa, đầu ra 3.5 cm là có thể, nhưng $p(3.5) = 0$ nhưng $p(\text{từ } 3 \text{ đến } 4) \neq 0$, tức xác suất thành phần riêng lẻ là 0. Thử nghĩ ra có giá trị xác suất rất nhỏ; theo thống kê, đó là 0.

Về hình thức, nếu X là một biến ngẫu nhiên liên tục, thì nó có hàm mật độ xác suất $f(x)$, do đó xác suất của nó rơi vào một khoảng nhất định, $[a, b]$, là tích phân $P[a \leq X \leq b] = \int_a^b f(x)dx$. Có thể, xác suất cho X để lấy bất kỳ giá trị đơn lẻ a , tức $a \leq X \leq a$ bằng 0, bởi vì một tích phân với giới hạn trên và dưới sát nhau luôn bằng 0.

Định nghĩa cho rằng phân phối xác suất liên tục phải có mật độ, hoặc tương đương, hàm phân phối tích lũy của nó là liên tục. Yêu cầu này mạnh hơn sự liên tục đơn giản của hàm phân phối tích lũy.

tích lũy, và có một lớp đặc biệt về phân phối, *phân phối đơn lẻ*, không liên tục hoặc không rời rạc cũng không phải là hỗn hợp của chúng. Chẳng hạn phân phối Cantor. Tuy nhiên những phân phối đơn lẻ như vậy chưa gặp phải trong thực tế.

Về thuật ngữ, một số người sử dụng thuật ngữ *phân phối liên tục* để biểu thị sự phân phối với hàm phân phối tích lũy liên tục. Do đó, định nghĩa của họ bao gồm cả phân phối hoàn toàn liên tục và phân phối đơn lẻ.

Theo quy ước, phân phối xác suất μ là liên tục nếu hàm phân phối tích lũy của nó $F(x) = \mu(-\infty, x] = \int_{-\infty}^x f(t)dt$. Các phân phối rời rạc và vài phân phối liên tục, chẳng hạn như phân phối Cantor, không chấp nhận mật độ này.

2.5.5.1. Một số thuộc tính của phân phối xác suất liên tục

Có vài thuộc tính:

- Sự phân phối xác suất của tổng của hai biến ngẫu nhiên độc lập là *tổng hợp*¹ của mỗi phân phối của chúng;
- Sự phân phối xác suất không phải là một không gian vector; chúng không là đóng theo tổ hợp tuyến tính, vì chúng không bảo toàn tính không âm hay toàn thể¹, nhưng chúng là đóng đối với tổ hợp lồi, do vậy tạo nên tập con lồi của không gian các hàm.

2.5.6. Định nghĩa Kolmogorov

Khi hình thành lí thuyết độ đo của lí thuyết xác suất, biến ngẫu nhiên được xác định như hàm độ đo X từ không gian xác suất (Ω, F, P) đến không gian (X, A) . *Phân phối xác suất của X* là độ đo về trước X^*P của X , là độ đo xác suất trên (X, A) thỏa mãn $X^*P = PX^{-1}$.

¹ Convolution: xoắn lại.

2.5.7. Tạo số ngẫu nhiên

Vấn đề về tần suất trong mô phỏng thống kê, như phương pháp Monte Carlo, là tạo ra các con số giả ngẫu nhiên được phân phối theo một cách nhất định. Hầu hết các thuật toán được dựa trên một máy phát số giả ngẫu nhiên, cho phép tạo ra các số $X \propto$ phân phối đồng đều trong đoạn $[0; 1]$. Các biến thể ngẫu nhiên λ sau đó được chuyển đổi thông qua một số thuật toán để tạo ra một biến thể ngẫu nhiên mới có phân phối xác suất yêu cầu.

2.5.8. Các ứng dụng

Khái niệm phân phối xác suất và các biến ngẫu nhiên được miêu tả trên nền toán học về lý thuyết xác suất, và khoa học thống kê. Sự lan truyền hoặc biến đổi gần như của bất kỳ giá trị nào đều có thể đo được, chẳng hạn chiều cao người, độ bền của kim loại, tăng trưởng doanh thu, lưu lượng giao thông... Gần như tất cả các phép đo được thực hiện với một số lỗi nội tại trong vật lý nhiều quá trình được mô tả theo xác suất, từ tính động học của khí đến mô tả cơ lượng tử của các hạt cơ bản... V những lý do này và nhiều lý do khác, các con số đơn giản thường không đủ để mô tả định lượng, trong khi phân phối xác suất thường thích hợp hơn.

Thí dụ cụ thể hơn về ứng dụng của phân tích thống kê như các mô hình ngôn ngữ nhanh và các mô hình ngôn ngữ thống kê khác được sử dụng trong xử lý ngôn ngữ tự nhiên để phân định xác suất xuất hiện của các từ và chuỗi từ đặc biệt, nhờ các phân phối xác suất.

2.5.9. Phân phối xác suất thông dụng

Dưới đây là danh sách một số phân phối xác suất phổ biến nhất, được nhóm lại theo quá trình mà chúng liên quan. Để hoàn thiện hơn, cần chia nhóm phân phối theo bản chất của đầu ra, như (i) rời rạc; (ii) liên tục; (iii) nhiều biến...

Các phân phối một biến dưới đây đều có hình chóp, tức có giá trị cụm xung quanh một điểm đơn. Trên thực tế, lực lượng thực tế quan sát được có thể bao quanh nhiều giá trị. Lực lượng như vậy có thể được mô hình hóa nhờ phân phối hỗn hợp.

2.5.9.1. Liên quan đến lượng số thực, có phân phối tuyến tính

Việc xuất hiện lỗi, độ lệch,... theo *phân phối chuẩn*, tức phân phối Gauss, đối với lượng đơn; thông dụng nhất là phân phối liên tục.

2.5.9.2. Liên quan đến lượng giá trị thực, dương, có phân phối theo hàm mũ

Các giá trị về giá cả, thu nhập, dân số,... theo phân phối:

- *Phân phối log chuẩn*, đối với lượng đơn, có log theo phân phối chuẩn;
- *Phân phối Pareto*, đối với lực lượng đơn, có log theo phân phối mũ.

2.5.9.3. Liên quan đến lượng giá trị thực, với giả thiết phân phối đồng đều trên một vùng

- *Phân phối đồng đều rời rạc*, đối với tập hợp hữu hạn các giá trị, chẳng hạn đầu ra về cái chết;
- *Phân phối đều liên tục*, đối với các giá trị phân phối liên tục.

2.5.9.4. Liên quan đến thử nghiệm Bernoulli, tức sự kiện có/không, với xác suất cho trước

- Các phân phối cơ bản (i) phân phối Bernoulli, đối với đầu ra của phép thử Bernoulli đơn, về thành công/ hay không thành công...; (ii) phân phối nhị thức, được số các xuất hiện dương, tức thành công, khi biết tổng số các xuất hiện thành công; (iii) Phân phối nhị thức âm, đối với quan sát kiểu nhị thức, nhưng lượng quan tâm là số các sai sót, thất bại trước khi có thành công đầu tiên; (iv) phổ biến hình học, đối với quan sát kiểu nhị thức, nhưng lượng quan tâm là số các sai sót, thất bại trước khi có thành công đầu tiên; đây là trường hợp đặc biệt của phân phối nhị thức âm;

- Các phân phối liên quan đến các lược đồ lấy mẫu trên một quần thể hữu hạn (i) phân phối siêu hình học, đối với số xuất hiện dương, chẳng hạn thành công, bầu bán..., và tổng số xuất hiện đã biết, dùng phương pháp lấy mẫu không thay thế; (ii) phân phối nhị thức beta, đối với số xuất hiện dương, chẳng hạn bầu, thành công... khi biết tổng số xuất hiện, lấy mẫu theo lược đồ Polya, tức trái với phương pháp lấy mẫu không thay thế.

2.5.9.5. Liên quan đến đầu ra phạm trù, các sự kiện với K đầu ra có thể, khi biết xác suất của mỗi đầu ra

- *Phân phối phạm trù*, đối với đầu ra phạm trù đơn, chẳng hạn trả lời có/không trong điều tra; là tổng quát hóa của phân phối Bernoulli;
- *Phân phối đa nhị thức*, đối với số của mỗi loại phạm trù đầu ra khi biết tổng đầu ra; là tổng quát hóa của phân phối nhị thức;
- *Phân phối siêu hình học đa biến*, tương tự như phân phối đa nhị thức, nhưng dùng phương pháp lấy mẫu không thay thế; là tổng quát hóa của phân phối siêu hình học.

2.5.9.6. Liên quan đến các sự kiện trong quá trình Poisson, tức các sự kiện xảy ra độc lập với tần suất đã biết

- *Phân phối Poisson*, đối với số xuất hiện của sự kiện kiểu Poisson trong một khoảng thời gian nhất định;
- *Phân phối mũ*, đối với thời gian trước sự kiện kiểu Poisson tiếp theo xảy ra;
- *Phân phối gamma*, đối với thời gian trước k sự kiện kiểu Poisson tiếp theo xảy ra;

2.5.9.7. Liên quan đến các giá trị tuyệt đối của vectơ với các thành phần phân phối chuẩn

- *Phân phối Rayleigh*, để phân phối để phân phối độ lớn của vectơ với các thành phần trực giao phân phối Gauss. Phân phối Rayleigh có trong các tín hiệu RF với các thành phần thực và ảo Gauss;

- *Phân phối Rice*, là tổng quát của phân phối Rayleigh cho vị trí có thành phần tín hiệu nền tĩnh. Nó có trong việc mờ dần Rice của tín hiệu radio truyền nhiều đường và trong ảnh MR có nhiễu ồn trên tín hiệu NMR khác không.

2.5.9.8. Liên quan đến lượng phân phối chuẩn, thao tác với tổng bình phương trong phép thử

- *Phân phối χ^2* , là phân phối của tổng các biến chuẩn bình phương, là có ích cho suy luận theo thay đổi của mẫu của các mẫu có phân phối chuẩn;
- *Phân phối Student*, hay T student, là phân phối của tỉ số biến chuẩn và căn bậc hai của biến χ^2 , có ích trong suy luận theo trung bình của mẫu chuẩn với các biến chưa biết;
- *Phân phối F*, là phân phối về tỉ số của hai biến χ^2 , có ích trong suy luận khi cần so sánh các biến hay cần đến hệ số tương quan bình phương.

2.5.10. Công dụng khi kết hợp phân phối tiên nghiệm trong suy luận Bayes

Một số phân phối xác suất:

- *Phân phối Beta*, đối với xác suất đơn, tức số thực giữa 0 và 1, kết hợp với phân phối Bernoulli và phân phối nhị thức;
- *Phân phối gamma*, đối với một tham số vô hướng, không âm; kết hợp với tham số tỉ lệ của phân phối Poisson hoặc phân phối mũ, độ chính xác, tức phương sai ngược, của phân phối chuẩn...
- *Phân phối Dirichlet*, đối với một véc tơ xác suất có tổng 1; kết hợp với phân phối phạm trù và phân phối đa nhị thức; là tổng quát của phân phối beta;
- *Phân phối Wishart*, đối với một ma trận xác định đối xứng không âm; kết hợp với nghịch đảo của ma trận hiệp phương sai của phân phối chuẩn nhiều biến; là tổng quát của phân phối gamma.

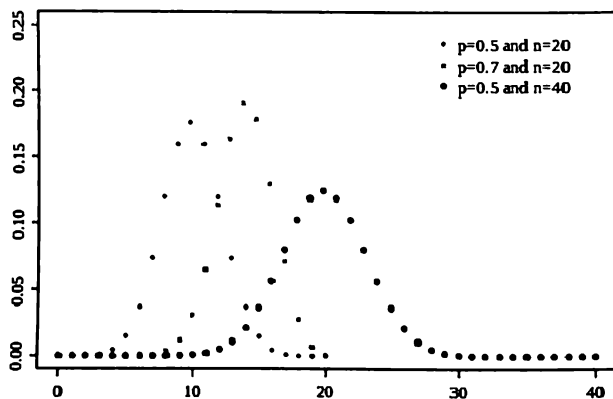
2.5.11. Một số phân phối một chiều quen thuộc

2.5.11.1. Phân phối nhị thức

Phép thử Bernoulli là phép thử chỉ quan tâm đến hai sự kiện, tức A và $\bar{A}$, với $p(A) = p$. Chẳng hạn tung đồng xu, được một trong hai mặt (i) sấp; (ii) ngửa. Biến ngẫu nhiên X có giá trị $X = 1$ khi xuất hiện A , 0 nếu ngược lại, tức xuất hiện $\bar{A}$. Khi đó X có phân phối Bernoulli, với tham số p , kí hiệu $X \in B(p)$.

Với n phép thử Bernoulli độc lập. Phép thử thứ i , với biến ngẫu nhiên $X_i \in B(p)$. Biến ngẫu nhiên X là số lần sự kiện A xuất hiện, trong n lần thử. Vậy $X = X_1 + X_2 + \dots + X_n$. X có phân phối nhị thức, kí hiệu $X \in B(n, p)$.

Xác suất trong n lần thử, xuất hiện k sự kiện A là $p_k = P(X = k) = C_n^k p^k q^{n-k}$, trong đó $q = 1 - p$, $C_n^k = \frac{n!}{k!(n-k)!}$.



Hình 2.22. Hàm mật độ xác suất, phân phối nhị thức.

Định nghĩa: Biến ngẫu nhiên X có phân phối nhị thức nếu xác suất xảy ra k lần trong n lần thử là $P(X = k) = C_n^k p^k q^{n-k}$.

2.5.11.2. Phân phối Poisson

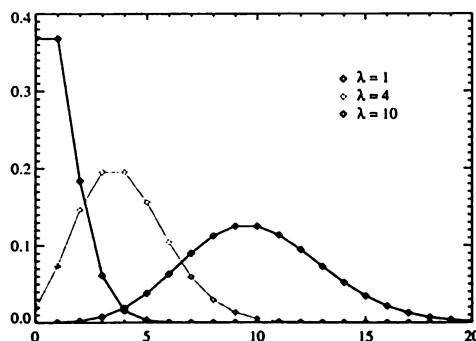
Một quá trình Poisson là một quá trình ngẫu nhiên mà trong đó sự kiện xảy ra liên tục và độc lập với nhau. Các ví dụ được mô

hình hóa tốt như các quá trình Poisson bao gồm (i) phân rã phóng xạ của các nguyên tử; (ii) các cuộc gọi điện thoại khi đến một tổng đài; (iii) truy cập vào một trang Web; (iv) lượng mưa...

Quá trình Poisson là một sưu tập $\{N(t): t \geq 0\}$ của các biến ngẫu nhiên, trong đó $N(t)$ là số các sự kiện đã xảy ra cho tới thời điểm t , bắt đầu từ thời điểm 0. Số lượng các sự kiện giữa thời gian a và thời gian b là $N(b) - N(a)$ và có phân phối Poisson. Mỗi thực hiện các quá trình $\{N(t)\}$ là một hàm bước nhảy, nguyên, không âm, là không giảm. Để trực quan, người ta xem nó như là một mô hình điểm trên $[0, \infty)$.

Bài toán cần đến phân phối Poisson có thể là: truy cập đến trang Web A xảy ra ngẫu nhiên, với tần suất trung bình λ lần truy cập trong một ngày. Một ngày chia thành n khoảng thời gian, sao cho trong mỗi khoảng có ít nhất một truy cập. Gọi X là việc xảy ra ít nhất một truy cập trong khoảng thời gian. Khả năng truy cập này là λ/n . Vậy $X \in B(n, \lambda/n)$. Theo cách tính trong phân phối nhị thức, $p(X = k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}$ khi $n \rightarrow \infty$.

Định nghĩa: Biến ngẫu nhiên X có phân phối Poisson¹ với tham số λ , kí hiệu $X \in P(\lambda)$, nếu $X = \{0, 1, \dots, n, \dots\}$ với xác suất $p_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.



Hình 2.23. Hàm mật độ¹ xác suất, phân phối Poisson, gồm các điểm

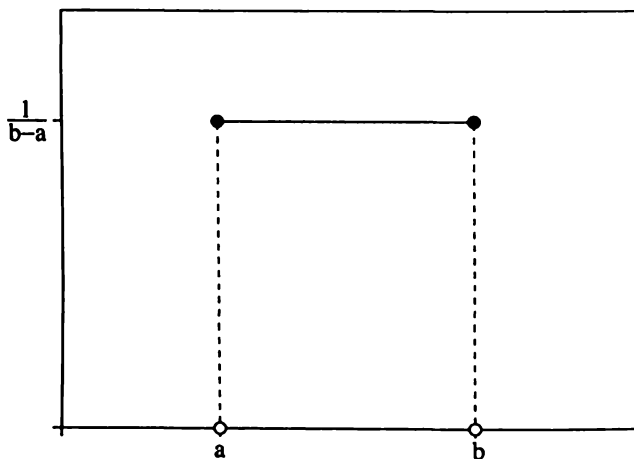
¹ Phân phối Poisson do Siméon Denis D. Poisson đưa ra năm 1837.

Trong định nghĩa, λ là trung bình số lần xuất hiện sự kiện quan tâm, trong khoảng thời gian, hay đơn vị tính nào đó.

2.5.11.3. Phân phối đều

Định nghĩa: Phân phối đều liên tục là một phân phối mà xác suất xảy ra như nhau cho mọi kết cục của biến ngẫu nhiên liên tục.

Phân phối đều liên tục đôi khi còn được gọi là phân phối hình chữ nhật và khi biểu diễn bằng hình vẽ sẽ có dạng hình chữ nhật. Hàm mật độ xác suất của một phân phối đều liên tục có dạng $f = 1/(b - a)$ đối với $x \in [a, b]$; ngược lại 0; trong đó x là biến ngẫu nhiên liên tục, a là giá trị cực tiểu, b là giá trị cực đại.



Hình 2.24. Hàm mật độ xác suất, phân phối đều liên tục

2.5.11.4. Phân phối chuẩn

Tự nhiên cho thấy một số sự kiện xảy ra theo quy luật, theo chuẩn. Chẳng hạn (i) cân nặng của trẻ sơ sinh; (ii) kết quả môn học; (iii) sai số đo... Trong kinh tế, nghiên cứu khoa học,

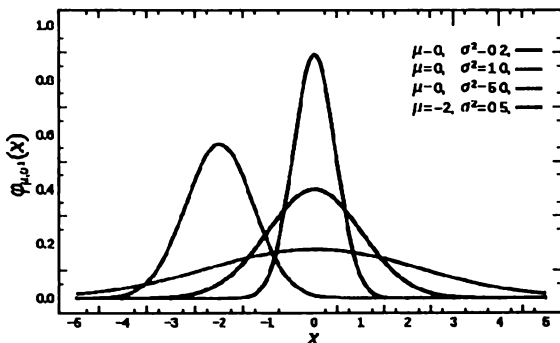
¹ Mass function: hàm khối.

một số sự kiện không theo phân phối chuẩn, nhưng xét toàn cục, về trung bình, người ta thấy vẫn có phân phối chuẩn.

Định nghĩa: Phân phối chuẩn, với trung bình μ và phương sai σ^2 , là phân phối có hàm mật độ xác suất có dạng $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Phân phối chuẩn¹, còn gọi là phân phối Gauss, là một phân phối xác suất quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số vị trí, tức giá trị trung bình μ , và tỉ lệ, tức phương sai σ^2 . Hàm mật độ xác suất là $f(x)$ hay $f(x; \mu, \sigma)$.

Định nghĩa: Phân phối chuẩn tắc² là phân phối chuẩn với giá trị trung bình bằng 0 và phương sai bằng 1.



Hình 2.25. Phân phối chuẩn và phân phối chuẩn tắc

Phân phối chuẩn còn được gọi là đường cong chuông, vì đồ thị của mật độ xác suất có dạng chuông.

2.5.11.5. Phân phối mũ

Nhiều đại lượng ngẫu nhiên tuân theo phân phối mũ, chẳng hạn (i) thời gian phục vụ khách hàng ăn, tức phục vụ đám đông;

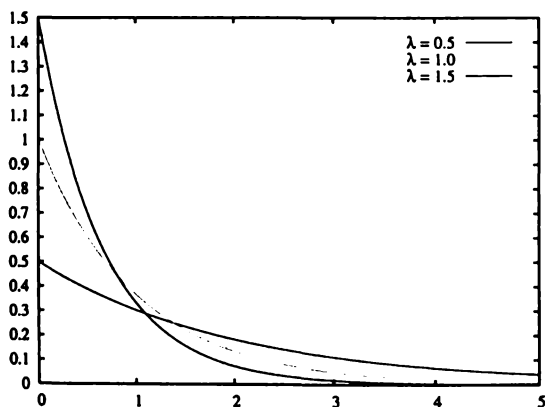
¹ Phân phối chuẩn do Abraham de Moivre đề xuất năm 1734.

² Standard normal distribution: phân phối chuẩn tắc.

(ii) thời gian nói chuyện trên điện thoại; (iii) thời gian sống củ thiết bị điện tử... Trong lí thuyết xác suất và thống kê, phân phối mũ là một lớp của các phân phối xác suất liên tục. Chúng thường được dùng để mô hình thời gian giữa các biến cố xảy ra theo tỉ lệ trung bình là hằng số.

Định nghĩa: Biến ngẫu nhiên X theo phân phối mũ khi hàm mật độ xác suất là $f(x) = \lambda e^{-\lambda x}$ nếu $x > 0$; ngược lại 0.

Trong định nghĩa, $\lambda > 0$; $f(x)$ có thể viết $f(x; \lambda)$.



Hình 2.26. Hàm mật độ xác suất, phân phối mũ

2.5.11.6. Phân phối Student

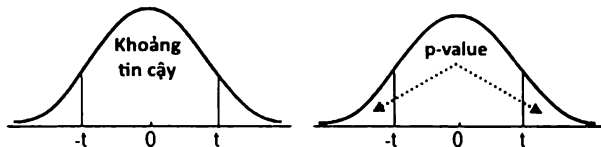
Định nghĩa: Biến ngẫu nhiên liên tục X có phân phối Student¹, hay phân phối t , với n bậc tự do, nếu X có hàm mật độ xác suất $f(x)$ được xác

$$\text{định theo } f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Trong định nghĩa, có hàm $\Gamma(x)$; được xác định $\Gamma(u) = \int_0^\infty e^{-x} x^{u-1} dx$. Về hàm này: $(u+1) = u \cdot \Gamma(u)$; $\Gamma(1) = 1$; $\Gamma(1/2) = \sqrt{\pi}$

¹ Phân phối Student, hay phân phối t , do William S. Gosset đề xuất năm 1908.

Phân phối Student cũng đối xứng quanh 0 như phân phối chuẩn, nhưng thấp hơn. Khi bậc tự do n càng lớn, phân phối Student càng gần đến phân phối chuẩn. Thực tế, khi bậc tự do lớn hơn 30, phân phối này được xem như phân phối chuẩn $N(0, n/(n-2))$.



Hình 2.27. Khi $X \in N(0; 1)$ là độc lập thống kê

$$\text{thì } t_k = \frac{X}{\sqrt{\frac{\chi_k^2}{k}}} \text{ có phân phối Student}$$

Để đánh giá độ khác biệt giữa hai nhóm, người ta thường sử dụng phương pháp kiểm định t . Kiểm định t là một trong những phương pháp đơn giản nhất trong thống kê học, vì có thể tính toán một cách thủ công, mà không cần đến máy tính hay phần mềm phân tích số liệu.

2.5.11.7. Phân phối χ^2

Trong thống kê, kiểm định χ^2 là họ các phương pháp kiểm định giả thuyết thống kê, mà kiểm định tuân theo phân phối χ^2 nếu đúng là có giả thuyết không. Có (i) kiểm định χ^2 Pearson; (ii) kiểm định χ^2 Yates; (iii) kiểm định χ^2 Mantel Haenszel...

Dạng thống kê kiểm định thông dụng nhất là $\chi^2 = \sum \frac{(o-e)^2}{e}$;

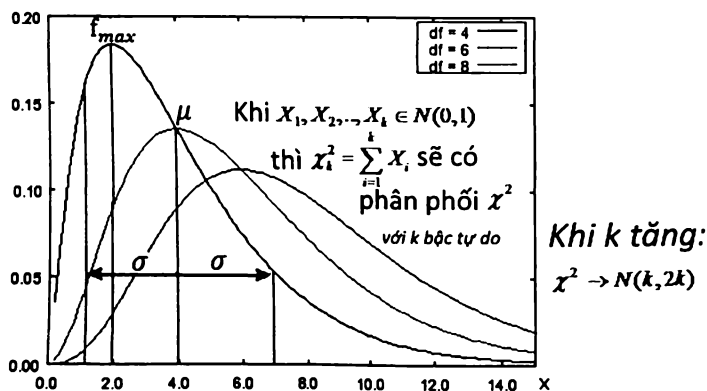
trong đó o là dữ liệu đo đạc, e là giá trị dự đoán chính xác. Phương pháp kiểm định χ^2 giúp tìm ra các xu hướng có thực từ một đo đạc chứa nhiều loạn ngẫu nhiên. Nó có thể áp dụng cho dữ liệu (i) lấy ngẫu nhiên từ một tập hợp; (ii) các đo đạc là độc lập với nhau... Phương pháp này cũng kiểm tra tính độc lập thống kê và mức độ khớp của dữ liệu. Chẳng hạn có nhóm người, người ta cần kiểm

định liên quan giữa (i) trình độ; (ii) nhóm tuổi, người ta dùng kiểm định χ^2 .

Định nghĩa: Biến ngẫu nhiên X có phân phối χ^2 bậc tự do n thì

hàm mật độ xác suất có dạng $f(x) = \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}}$ khi $x > 0$; ngược lại 0.

Phân phối χ^2 là phân phối lệch trái. Nếu $X_1, \dots, X_k$ là biến ngẫu nhiên độc lập, có phân phối chuẩn tắc $N(0; 1)$, thì $\chi_k^2 = \sum_{i=1}^k X_i^2$, tức theo phân phối χ^2 . Phân phối χ^2 có tính chất (i) lệch trái; khi bậc tự do tăng thì nó tiến dần về phân phối chuẩn $N(k, 2k)$; (ii) tổng các biến có phân phối χ^2 cũng có phân phối χ^2 với bậc tự do bằng tổng các bậc tự do thành phần, tức $\chi_{k_1+k_2}^2 = \chi_{k_1}^2 + \chi_{k_2}^2$.



Hình 2.28. Phân phối χ^2

2.5.12. Một số đặc trưng của đại lượng ngẫu nhiên

2.5.12.1. Trung bình

Định nghĩa: Trung bình cộng có trọng số, hay số bình quân g quyền, của một tập là giá trị trung bình cộng có phản ánh tầm quan trọng của các phần tử, hay giá trị quan sát trong tập đó.

Mỗi một giá trị quan sát sẽ được gán một trọng số. Công thức tính trung bình cộng có trọng số là $\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$, trong đó $x_1, x_2, \dots$

x_n là các phần tử trong tập, và $w_1, w_2, \dots, w_n$ là các trọng số tương ứng của từng phần tử, $i = 1..n$. Trong thống kê, trung bình cộng có trọng số hay được dùng để tính toán các chỉ số.

2.5.12.2. Điểm giữa

Định nghĩa: Trong lý thuyết xác suất, thống kê, trung vị được xem như giá trị số tách nửa cao của mẫu, của lực lượng, hay của phân phối xác suất, với nửa thấp hơn.

Điểm giữa của danh sách hữu hạn có thể xuất hiện trong dãy quan sát như giá trị ở giữa trong số các giá trị cao và thấp. Trong các dữ liệu mẫu, hay trong toàn lực lượng, không có phần tử với giá trị khớp với điểm giữa, như số phần tử là chẵn, và nếu có một phần tử, thì có thể điểm giữa không trùng với phần tử của mẫu. Ngoài khái niệm điểm giữa, khái niệm liên quan đến phần tử của mẫu là tâm đại diện. Chẳng hạn nếu $a < b < c$, thì điểm giữa của danh sách $\{a, b, c\}$ là b , và nếu $a < b < c < d$, thì điểm giữa của danh sách $\{a, b, c, d\}$ là trung bình của b và c , tức $(b + c)/2$. Cho dãy số, người ta tính được điểm giữa (i) khi số các phần tử là lẻ: Cho danh sách 1, 5, 2, 8, 7. Sắp xếp tăng dần được 1, 2, 5, 7, 8. Khi đó 5 là điểm giữa vì nó là quan sát ở giữa trong danh sách đã xếp; (ii) khi số các phần tử là chẵn: Cho danh sách 1, 5, 2, 8, 7, 2. Danh sách đã xếp là 1, 2, 2, 5, 7, 8. Trung bình của phần tử giữa là $(2 + 5)/2 = 3.5$. Do vậy điểm giữa là 3.5 vì nó là trung bình của các quan sát giữa trong danh sách đã xếp.

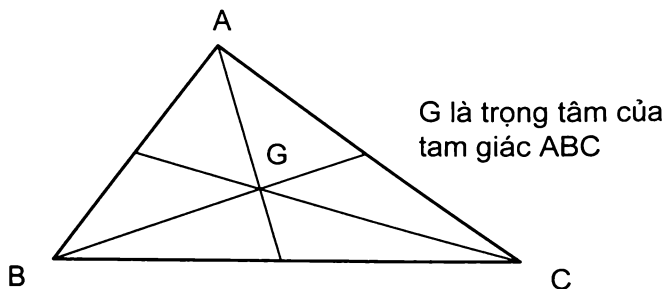
Điểm giữa có thể được dùng như thước đo vị trí khi phân phối là lệch, khi các giá trị biên chưa biết, hay khi cần giảm mức quan trọng đối với lề. Nhược điểm của điểm giữa là sự khó khăn quản lý nó về mặt lý thuyết. Điểm giữa của biến x được kí hiệu là $\tilde{x}$ hay $\mu_{1/2}(x)$.

2.5.12.3. Tâm đại diện

Định nghĩa: Tâm đại diện¹ là các đối tượng đại diện của bộ dữ liệu hay của cụm, có sai khác trung bình so với tất cả các đối tượng khác trong cụm là nhỏ nhất.

Tâm đại diện là tương tự về khái niệm với trung bình trọng tâm, nhưng tâm đại diện luôn luôn là các phần tử của bộ dữ liệu. Tâm đại diện thường được sử dụng với dữ liệu khi người ta không xác định được trung bình hay trọng tâm, như các quỹ đạo 3D, trong ngữ cảnh thể hiện gen.

2.5.12.4. Trọng tâm



Hình 2.29. Trọng tâm của một tam giác

Định nghĩa: Trong hình học, trọng tâm², tâm hình học, hay điểm giữa của hình phẳng hay hình hai chiều.

Trọng tâm là giao của tất cả các đường thẳng mà chia hình thành hai phần có mômen bằng nhau. Nói theo cách phi hình thức, đây là trung bình của tất cả các điểm trên mặt phẳng. Định nghĩa này mở rộng cho đối tượng X trong không gian n chiều: trọng tâm là giao của tất cả các siêu phẳng cho phép chia X thành hai phần có mômen như nhau.

¹ Medoid: tâm đại diện.

² Centroid: trọng tâm.

2.5.12.5. Phần trăm

Định nghĩa: Phần trăm¹ là giá trị của biến mà dưới đó có lượng phần trăm của các quan sát rơi vào.

Giải thích qua thí dụ, phần trăm thứ 20 là giá trị, hay tỉ số, mà 20% các quan sát được thấy. Thuật ngữ *phần trăm* và thuật ngữ liên quan như hạng phần trăm thường được sử dụng trong thống kê mô tả, cũng như trong thông báo về tỉ số từ các thử nghiệm theo chuẩn.

Phần trăm thứ 25 được gọi là *Quartile* đầu tiên, kí hiệu Q_1 ; phần trăm thứ 50 được gọi là trung vị, hay *Quartile* thứ hai, Q_2 ; và phần trăm thứ 75 là *Quartile* thứ ba, Q_3 .

2.6. Góc nhìn về phân tích dữ liệu

2.6.1. Ma trận dữ liệu

Xét ma trận dữ liệu D , d cột ứng với các thuộc tính quan tâm.

$$D = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Hình 2.30. Ma trận dữ liệu

Kí hiệu x_i là dòng thứ i , $x_i = (x_{i1}, \dots, x_{id})$; kí hiệu X_j là cột thứ j , $X_j = (x_{1j}, \dots, x_{nj})$. Tùy theo lĩnh vực ứng dụng, mà dòng được tham chiếu như *thực thể, thể hiện, thí dụ, bản ghi, đối tượng, vector đặc trưng, bộ*; cột được tham chiếu như *thuộc tính, đặc trưng, chiều*.

¹ Percentile: phần trăm.



Hình 2.31. Hoa diên vĩ

Số *d* gọi là *chiều* của dữ liệu. Phân tích thuộc tính đơn gọi là *phân tích một biến*; phân tích đồng thời nhiều chiều gọi là *phân tích nhiều biến*.

Thí dụ trích trong bộ dữ liệu về *hoa diên vĩ*¹. Dữ liệu tạo nên ma trận 150*5. Thực thể là hoa diên vĩ, các thuật toán là (i) độ dài của lá đài; (ii) độ rộng của lá đài; (iii) độ dài của cánh hoa; (iv) độ rộng của cánh hoa; (v) lớp hoa.

2.6.2. Thuộc tính

Các thuộc tính cho phép phân lớp thành hai kiểu chính, the lĩnh vực, tức theo kiểu dữ liệu mà đối tượng nhận.

2.6.2.1. Thuộc tính số

Định nghĩa: Thuộc tính số là thuộc tính nhận giá trị số thực hay số nguyên.

Các thuộc tính số có thể (i) hữu hạn; (ii) vô hạn. Khi đối tượng nhận hữu hạn giá trị trong tập, người ta có dữ liệu *rời rạc*. Khi đối tượng nhận vô hạn giá trị thực, nó *liên tục*. Các thuộc tính số được tách thành hai dạng:

¹ Họ diên vĩ (iris): Các loài trong họ này là cây thường xanh, với thân hành lá thân củ hay thân rễ. Các loài cây này nói chung có thân mọc thẳng và các lá có hình dáng giống lá cỏ, với nếp gấp trung tâm sắc nhọn.

ng 2. Dữ liệu để phân tích thống kê và khai phá

Bảng 2.9. Bảng dữ liệu về hoa diên vĩ

	Độ dài cánh hoa	Độ rộng cánh hoa	Độ dài lá đài hoa	Độ rộng lá đài hoa	Lớp
	X_1	X_2	X_3	X_4	X_5
x_1	5.9	3.0	4.2	1.5	Hoa diên vĩ nhiều màu
x_2	6.9	3.1	4.9	1.5	Hoa diên vĩ nhiều màu
x_3	6.6	2.9	4.6	1.3	Hoa diên vĩ nhiều màu
x_4	4.6	3.2	1.4	0.2	Hoa diên vĩ mượt
x_5	6.0	2.2	4.0	1.0	Hoa diên vĩ nhiều màu
x_6	4.7	3.2	1.3	0.2	Hoa diên vĩ mượt
x_7	6.5	3.0	5.8	2.2	Hoa diên vĩ nhiều màu
x_8	5.8	2.7	5.1	1.9	Hoa diên vĩ có hoa
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{149}	7.7	3.8	6.7	2.2	Hoa diên vĩ có hoa
x_{150}	5.1	3.4	1.5	0.2	Hoa diên vĩ mượt

Thang đo khoảng. Loại thuộc tính này chú trọng sự sai Chằng hạn hôm nay 20°C, mai là 5°C, chú trọng đ 15°C;

Thang đo tỉ lệ. Loại thuộc tính này chú trọng cả độ lệc tỉ lệ hơn kém. Chằng hạn A là 20 tuổi, B là 5; họ hơn 15 và gấp 4 lần.

Các thuộc tính phạm trù

Định nghĩa: Thuộc tính phạm trù là loại thuộc tính lấy , trong tập hợp và các kí hiệu.

ác thuộc tính phạm trù được chia làm hai loại:

Theo tên¹. Các giá trị thuộc tính trong lĩnh vực khô thứ tự. Chằng hạn tên = {ấm, chén, ...};

Có thứ tự¹. Các giá trị thuộc tính là có thứ tự. Chẳn điểm = {4; 5; 6; ...}.

uinal: theo tên.

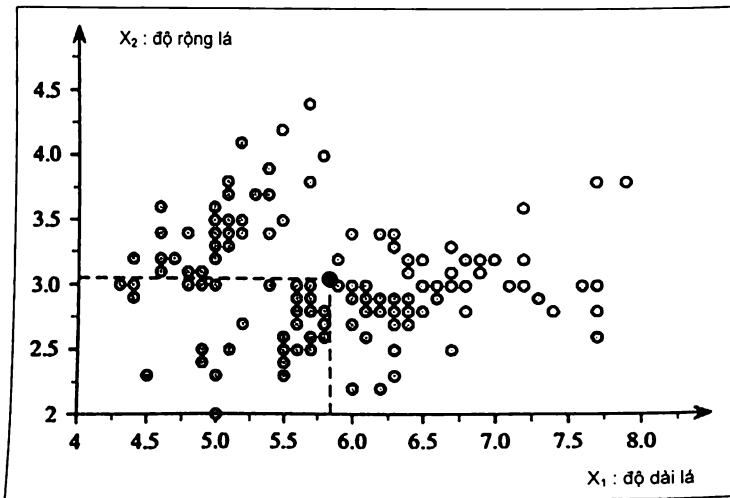
2.6.3. Khía cạnh đại số và góc nhìn hình học về dữ liệu

Nếu d thuộc tính trong ma trận D là số, dòng i được thể hiện qua $x_i = (x_{i1}, \dots, x_{id})$, tức điểm d chiều. Mỗi dòng có thể xem như vector cột d chiều, tức các vector giả thiết là vector cột, thì

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix} = (x_{i1}, x_{i2}, \dots, x_{id})^T \in R^d; \text{ trong đó } T \text{ là phép chuyển vị ma trận.}$$

Không gian tọa độ Decac được xác định qua d vector đơn vị, gọi là vector cơ sở chuẩn, theo mỗi trục. Vector cơ sở chuẩn thứ j là $e_j = (0, \dots, 0, 1, 0, \dots, 0)^T$. Bất kì vector nào trong R^d đều viết được qua tổ hợp tuyến tính của các vector cơ sở chuẩn; $x_i = x_{i1}e_1 + \dots + x_{id}e_d$.

Với dữ liệu trong bảng trên, chiếu dữ liệu lên hai chiều đầu tiên, thì mỗi dòng được xem như một điểm trong không gian 2 chiều. Chẳng hạn $x_1 = (5.9, 3.0, 4.2, 1.5, \text{diên vĩ nhiều màu})$. Hình minh họa đối với $n = 150$.



Hình 2.32. Đồ thị các điểm; điểm trung bình ở giữa

1 Ordinal: có thứ tự.

$$x_1 = 5.9e_1 + 3e_2 + 4.2e_3 = 5.9 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 4.2 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 5.9 \\ 3.0 \\ 4.2 \end{pmatrix}$$

Mỗi cột, hay thuộc tính được xem như vector trong không gian

$$n \text{ chiều } X_j = \begin{pmatrix} x_{1j} \\ \dots \\ x_{nj} \end{pmatrix}; D = \left(\begin{array}{c|c|c} x_{11} & x_{12} & x_{1d} \\ x_{21} & x_{22} & x_{2d} \\ \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{nd} \end{array} \right) = \begin{pmatrix} -x_1^T & - \\ -x_2^T & - \\ \dots & \dots \\ -x_n^T & - \end{pmatrix} = \left(\begin{array}{c|c|c} | & | & | \\ X_1 & \dots & X_d \\ | & | & | \end{array} \right).$$

2.6.3.1. Khoảng cách và góc

Xét các thể hiện dữ liệu và thuộc tính như các vector; bộ dữ liệu như ma trận. Cho $a, b \in \mathbb{R}^m$, là vector m chiều; $a = (a_1, \dots, a_m)$, $b = (b_1, \dots, b_m)$.

Định nghĩa: Tích chấm¹ giữa a và b được xác định như giá trị vô hướng $A^T b$.

Theo định nghĩa, tích chấm là $a^T b = (a_1 \ a_2 \dots a_m)^* (b_1 \ b_2 \dots b_m)^T = \sum_{i=1}^m a_i b_i$.

Định nghĩa: Độ dài, hay chuẩn Oclit của vector a được xác định $\|a\|$.

$$\|a\| = \sqrt{a^T a} = \sqrt{a_1^2 + \dots + a_n^2} = \sqrt{\sum_{i=1}^n a_i^2}$$

Vector đơn vị theo phương của a là $u = a / \|a\| = (1 / \|a\|) * a$. Xác định vector u , với $\|u\| = 1$, là vector chuẩn hóa, được sử dụng thay a trong các nhiệm vụ phân tích.

Chuẩn Oclit là trường hợp đặc biệt của chuẩn, gọi là chuẩn L_p . Nó được xác định với $p \neq 0$. Chuẩn Oclit là trường hợp $p = 2$.

¹ Dot product: tích chấm.

$$\|a\|_p = (|a_p|^p + \dots + |a_m|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^m |a_i|^p\right)^{\frac{1}{p}}$$

Định nghĩa: Khoảng cách Oclit giữa a và b được xác định $d(a, b)$.

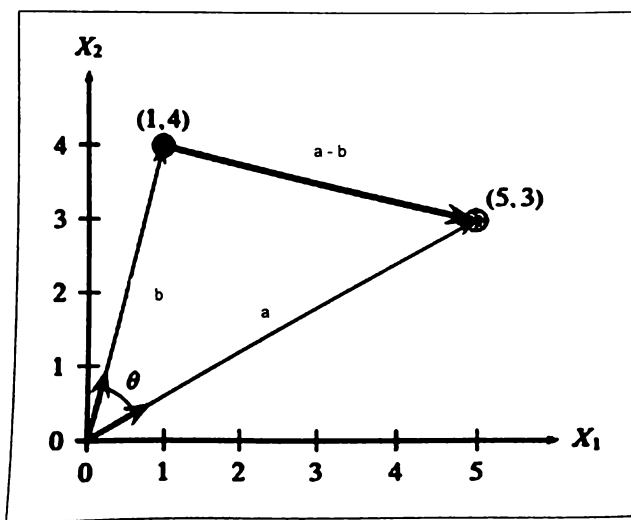
$$d(a, b) = \|a - b\| = \sqrt{(a - b)^T (a - b)} = \sqrt{\sum_{i=1}^m (a_i - b_i)^2}$$

Độ dài vector chính là khoảng cách của nó đến vector 0, tức $\|a\| = \|a - 0\| = d(a, 0)$. Trong chuẩn p , $d_p(a, b) = \|a - b\|_p$.

Định nghĩa: Góc θ giữa hai vector a và b là tương tự cosin giữa a và b , $\cos\theta = \frac{a^T b}{\|a\| \|b\|} = (a / \|a\|)^T (b / \|b\|)$.

Do vậy cosin của góc giữa a và b là tích chấm của vector đơn vị $a / \|a\|$ và $b / \|b\|$. Bất đẳng thức Cauchy - Schwartz khẳng định rằng đối với bất kỳ a, b : $|a^T b| \leq \|a\| * \|b\|$.

Định nghĩa: Hai vector a, b là trực giao nếu $a^T b = 0$, tức $\cos\theta = 0$, thì góc giữa chúng là 90° . Lúc này người ta nói chúng không giống nhau.



Hình 2.33. Khoảng cách và góc giữa hai vector

Về khoảng cách và góc, xét hai vector $a = (5; 3)$, $b = (1; 4)$. Khoảng cách được tính: $d(a, b) = \sqrt{(5-1)^2 + (3-4)^2} = \sqrt{17} = 4.12$; đồng thời $a - b = (5; 3) - (1; 4) = (4; -1)$, và $\|a - b\| = \sqrt{4^2 + (-1)^2} = \sqrt{17}$.

Vector đơn vị theo hướng của vector a là $u_a = a/\|a\| = \frac{1}{\sqrt{5^2 + 3^2}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \frac{1}{\sqrt{34}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.86 \\ 0.51 \end{pmatrix}$. Vector đơn vị theo hướng b là $u_b = (0.24, 0.97)$.

Góc θ giữa a và b là $\cos \theta = \frac{\begin{pmatrix} 5 \\ 3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 4 \end{pmatrix}}{\sqrt{5^2 + 3^2} \cdot \sqrt{1^2 + 4^2}} = \frac{17}{\sqrt{34} \cdot \sqrt{17}} = \frac{1}{\sqrt{2}}$, tức $\theta = 45^\circ$.

Xét chuẩn L_p đối với a , $p = 3$: $\|a\|_3 = (5^3 + 3^3)^{1/3} = (152)^{1/3} = 5.34$. Khoảng cách theo chuẩn L_p , với $p = 3$ là $\|a - b\|_3 = \|(4, -1)^T\|_3 = (4^3 + |-1|^3)^{1/3} = 4.02$

2.6.3.2. Trung bình và biến thiên toàn bộ

Khái niệm về trung bình:

Định nghĩa: Trung bình của ma trận D là vector như trung bình của tất cả các điểm $t_b(D) = \mu = (x_1 + \dots + x_n)/n$.

Khái niệm về biến thiên:

Định nghĩa: Biến thiên toàn bộ của ma trận D là khoảng cách bình phương trung bình của các điểm đối với trung bình: $\text{var}(D) = (d(x_1, \mu)^2 + \dots + d(x_n, \mu)^2)^{1/2} / n = (\|x_1 - \mu\|^2 + \dots + \|x_n - \mu\|^2) / n$.

Khái niệm về ma trận dữ liệu:

Định nghĩa: Ma trận dữ liệu giữa¹ thu được từ ma trận ban đầu bằng cách trừ các điểm một lượng bằng trung bình. $Z = D - 1 \cdot \mu^T = (x_1^T, \dots, x_n^T) - (\mu^T, \dots, \mu^T) = (x_1^T - \mu^T, \dots, x_n^T - \mu^T) = (z_1^T, \dots, z_n^T)$.

¹ Centered data matrix: ma trận dữ liệu giữa, hướng tâm.

$$\|a\|_p = (|a_p|^p + \dots + |a_m|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^m |a_i|^p\right)^{\frac{1}{p}}$$

Định nghĩa: Khoảng cách Oclit giữa a và b được xác định $d(a, b)$.

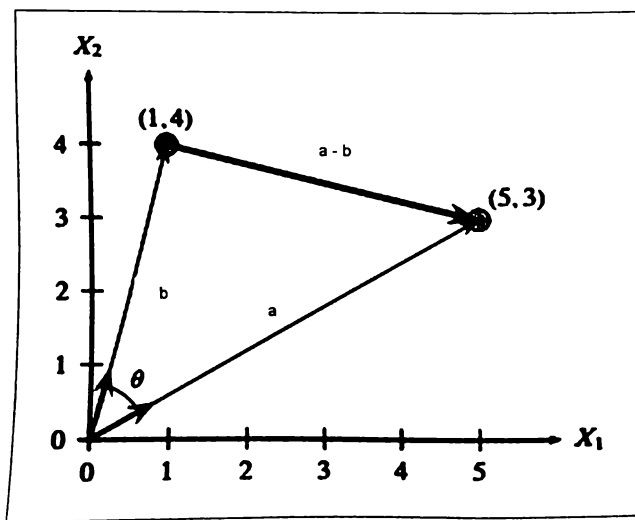
$$d(a, b) = \|a - b\| = \sqrt{(a - b)^T (a - b)} = \sqrt{\sum_{i=1}^m (a_i - b_i)^2}$$

Độ dài vector chính là khoảng cách của nó đến vector 0, tức $\|a\| = \|a - 0\| = d(a, 0)$. Trong chuẩn p , $d_p(a, b) = \|a - b\|_p$.

Định nghĩa: Góc θ giữa hai vector a và b là tương tự cosin giữa a , $\cos\theta = A^T b / \|a\| \|b\| = (a / \|a\|)^T (b / \|b\|)$.

Do vậy cosin của góc giữa a và b là tích chấm của vector đơn vị $a / \|a\|$ và $b / \|b\|$. Bất đẳng thức Cauchy - Schwartz khẳng định rằng đối với bất kỳ a, b : $|a^T b| \leq \|a\| * \|b\|$.

Định nghĩa: Hai vector a, b là trực giao nếu $a^T b = 0$, tức $\cos\theta = 0$, thì góc giữa chúng là 90° . Lúc này người ta nói chúng không giống nhau.



Hình 2.33. Khoảng cách và góc giữa hai vector

Về khoảng cách và góc, xét hai vectơ $a = (5; 3)$, $b = (1; 4)$. Khoảng cách được tính: $d(a, b) = \sqrt{(5-1)^2 + (3-4)^2} = \sqrt{17} = 4.12$; đồng thời $a - b = (5; 3) - (1; 4) = (4; -1)$, và $\|a - b\| = \sqrt{4^2 + (-1)^2} = \sqrt{17}$.

Vectơ đơn vị theo hướng của vectơ a là $u_a = a/\|a\| = \frac{1}{\sqrt{5^2 + 3^2}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \frac{1}{\sqrt{34}} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.86 \\ 0.51 \end{pmatrix}$. Vectơ đơn vị theo hướng b là $u_b = (0.24, 0.97)$.

Góc θ giữa a và b là $\cos \theta = \frac{\begin{pmatrix} 5 \\ 3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 4 \end{pmatrix}}{\sqrt{5^2 + 3^2} \cdot \sqrt{1^2 + 4^2}} = \frac{17}{\sqrt{34} \cdot \sqrt{17}} = \frac{1}{\sqrt{2}}$, tức $\theta = 45^\circ$.

Xét chuẩn L_p đối với a , $p = 3$: $\|a\|_3 = (5^3 + 3^3)^{1/3} = (152)^{1/3} = 5.34$. Khoảng cách theo chuẩn L_p , với $p = 3$ là $\|a - b\|_3 = \|(4, -1)^T\| = (4^3 + |-1|^3)^{1/3} = 4.02$

2.6.3.2. Trung bình và biến thiên toàn bộ

Khái niệm về trung bình:

Định nghĩa: Trung bình của ma trận D là vectơ như trung bình của tất cả các điểm $tb(D) = \mu = (x_1 + \dots + x_n)/n$.

Khái niệm về biến thiên:

Định nghĩa: Biến thiên toàn bộ của ma trận D là khoảng cách bình phương trung bình của các điểm đối với trung bình: $var(D) = (d(x_1, \mu)^2 + \dots + d(x_n, \mu)^2)^{1/2} / n = (\|x_1 - \mu\|^2 + \dots + \|x_n - \mu\|^2)^{1/2} / n$.

Khái niệm về ma trận dữ liệu:

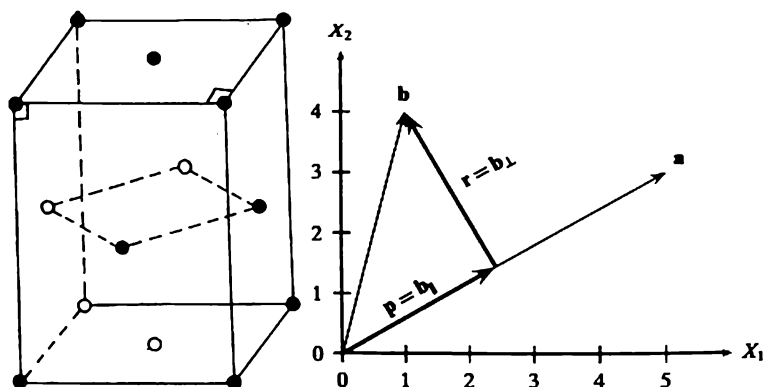
Định nghĩa: Ma trận dữ liệu giữa¹ thu được từ ma trận ban đầu bằng cách trừ các điểm một lượng bằng trung bình. $Z = D - 1 \cdot \mu^T = (x_1^T, \dots, x_n^T) - (\mu^T, \dots, \mu^T) = (x_1^T - \mu^T, \dots, x_n^T - \mu^T) = (z_1^T, \dots, z_n^T)$.

¹ Centered data matrix: ma trận dữ liệu giữa, hướng tâm.

2.6.3.3. Chiếu trực giao

Trong khai phá dữ liệu, người ta thường chiếu một điểm hay vector lên vector khác; chẳng hạn để biết điểm mới sau khi thay đổi vector cơ sở.

Định nghĩa: Cho $a, b \in R^m$, là hai vector, phân rã trực giao của vector b theo hướng của vector a là $b = p + r = b_{\parallel} + b_{\perp}$; trong đó p song song với a ; r là trực giao, hay thẳng góc với a .



Hình 2.34. Chiếu trực giao

Theo định nghĩa trên, $r = b - p$ cho biết khoảng cách giữa a và b , thường được diễn giải như vector lỗi giữa điểm b và p .

2.6.3.4. Độc lập tuyến tính và chiều

Cho ma trận $D = (x_1 \ x_2 \ \dots \ x_d) = (X_1 \ X_2 \ \dots \ X_d)$, người ta thường quan tâm đến tổ hợp tuyến tính của hàng hay cột. Các tổ hợp tuyến tính khác nhau của d thuộc tính cho ra các thuộc tính dẫn xuất, có vai trò như trích đặc trưng hay giảm chiều.

Định nghĩa: Cho tập các vector $v_1, \dots, v_k$ trong không gian vector m chiều; tổ hợp tuyến tính của chúng là $c_1 v_1 + \dots + c_k v_k$; trong đó c_k là giá trị vô hướng.

Tập các tổ hợp tuyến tính của k vector là *nhập*¹, kí hiệu $\text{span}(v_1, \dots, v_k)$. Nó là không gian con của R^m . Khi $\text{span}(v_1, \dots, v_k) = R^m$, thì $v_1, \dots, v_k$ là *tập nhập* của R^m .

Không gian dòng và cột: đối với ma trận D , hai thứ liên quan là (i) không gian cột; (ii) không gian dòng. Người ta xác định không gian cột $\text{cột}(D) = \text{span}(X_1, \dots, X_d)$; không gian dòng $\text{dòng}(D) = \text{span}(x_1, \dots, x_n)$.

Độc lập tuyến tính: khi các vector phụ thuộc, có các hằng số c_i để $c_1v_1 + \dots + c_kv_k = 0$. Khi $\{v_i\}$ không phụ thuộc tuyến tính, nó là *độc lập tuyến tính*. Điều này có nghĩa chỉ khi $c_i = 0$ thì $c_1v_1 + \dots + c_kv_k = 0$.

Định nghĩa: Các vector $v_1, \dots, v_k$ là *phụ thuộc tuyến tính*, nếu ít nhất có vector được biểu diễn như tổ hợp tuyến tính của các vector khác.

Chiều và hạng: cơ sở chuẩn đối với R^m là các vector $e_1 = (1, \dots, 0)$, $\dots$, $e_m = (0, \dots, 1)$. Hai cơ sở bất kì của S cần có cùng số vector; số này là *chiều* của S , kí hiệu *dim* (S). Do S là không gian con, $\text{dim}(S) \leq m$. Đối với ma trận bất kì, chiều của dòng và cột là như nhau; được gọi là *hạng*² của ma trận. Với ma trận $D \in R^{n \times d}$, $\text{rank}(D) \leq \min(n, d)$.

Định nghĩa: Cơ sở đối với không gian con S của R^m , là các vector $v_1, \dots, v_k$, *độc lập tuyến tính*, và là *nhập* của S , tức $\text{span}(v_1, \dots, v_k) = S$.

Khái niệm về cơ sở trực giao:

Định nghĩa: Cơ sở trực giao đối với S là cơ sở mà các vector đôi một trực giao. Nếu các vector là vector đơn vị, cơ sở này gọi là cơ sở chuẩn trực giao đối với S .

2.6.4. Góc nhìn xác suất đối với dữ liệu

2.6.4.1. Khái niệm

Góc nhìn xác suất của dữ liệu giả sử mỗi thuộc tính số X là *biến ngẫu nhiên*, được xác định như một hàm, gán giá trị thực cho mỗi kết quả của thực nghiệm. Theo cách hình thức, X là hàm X :

¹ Span: nhập.

² Rank: hạng.

$O \rightarrow R$, trong đó O là miền xác định của X , là tập các kết quả thực nghiệm, cũng được gọi là *không gian mẫu*; R là *hạng* của X , là tập các số thực.

Định nghĩa: X là biến ngẫu nhiên rời rạc nếu X chỉ lấy hữu hạn hoặc vô hạn đếm được các giá trị trong hạng của nó. X là biến ngẫu nhiên liên tục nếu nó nhận giá trị bất kì trong hạng của nó.

Xét độ dài lá đài hoa X_1 trong cơ sở dữ liệu về hoa diên vĩ trong bảng sau. Tất cả $n = 150$ giá trị của thuộc tính này cho trong bảng, trong hạng $[4.3, 7.9]$.

Mặc định thuộc tính X_1 là biến ngẫu nhiên liên tục, cho qu hàm $X_1(v) = v$, do các kết quả, về giá trị độ dài, đều là số. Nếu phân biệt hai loại hoa (i) độ dài cánh đài hoa ngắn; (ii) độ dài cánh đài hoa dài, với ngưỡng 7 cm, người ta có thể xác định biến ngẫu nhiên rời rạc A : $A(v) = 0$ nếu $v < 7$; $A(v) = 1$ nếu $v \geq 7$. Khi đó, miền xác định của A là $[4.3, 7.9]$ và hạng là $\{0, 1\}$. Do vậy A là biến ngẫu nhiên rời rạc khác không chỉ với giá trị rời rạc 0 và 1.

Bảng 2.10. Bộ dữ liệu về hoa diên vĩ, độ dài lá đài, theo cm

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Hàm xác suất toàn thể: xác suất mà cần thỏa mãn $f(x) \geq 0$ và $\sum f(x) = 1$.

Định nghĩa: Nếu X rời rạc, hàm xác suất toàn thể là $f(x) = P(X = x)$ với $x \in R$.

¹ Probability Mass Function: hàm mật độ xác suất.

Thí dụ về phân phối nhị thức và Bernoulli: trong thí dụ trước, A được xác định là biến ngẫu nhiên rời rạc. Trong bảng dữ liệu chỉ có 13 hoa có độ dài ít nhất 7 cm. Người ta ước lượng hàm xác suất toàn thể của A : $f(1) = P(A = 1) = 13/150 = 0.087 = p$; $f(0) = P(A = 0) = 137/150 = 0.913 = 1 - p$.

Người ta nói A có *phân phối Bernoulli* với tham số $p \in [0, 1]$, cho biết xác suất *thành công*, tức xác suất lấy được bông hoa với cánh dài dài, từ tập tất cả các điểm. Và $1 - p$ là xác suất *thất bại*, tức không lấy được hoa có cánh dài dài.

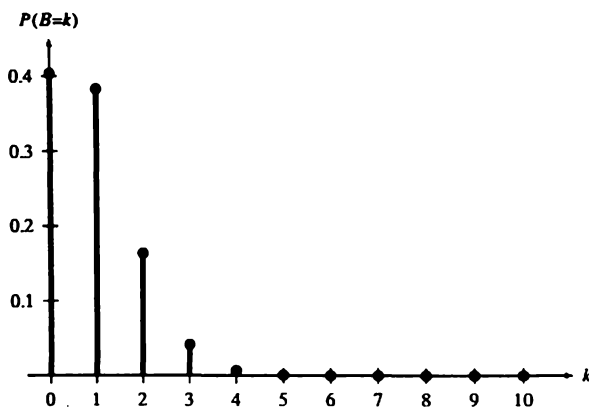
Lấy biến ngẫu nhiên rời rạc B khác, cho biết k bông hoa cánh dài dài trong số m phép thử Bernoulli độc lập, với xác suất thành công p . B nhận giá trị rời rạc $[0, m]$, và hàm xác suất toàn thể có dạng *phân phối nhị thức* $f(k) = P(B = k) = \binom{m}{k} p^k (1-p)^{m-k}$. Công

thức có nghĩa: có $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ cách để lấy k bông hoa cánh dài dài trong m lần thử. Đối với mỗi lựa chọn, xác suất tổng của k thành công là p^k , và xác suất tổng $m - k$ thất bại là $(1-p)^{m-k}$. Chẳng hạn $p = 0.087$ như trên, xác suất để chính xác $k = 2$ bông hoa cánh dài dài, trong $m = 10$ phép thử là $f(2) = P(B = 2) = \binom{10}{2} * 0.087^2 * 0.913^8 = 0.164$.

Hình vẽ cho thấy hàm xác suất toàn thể đối với các giá trị k , với $m = 10$. Do p là nhỏ, xác suất của k thành công trong số ít phép thử giảm nhanh khi tăng k , dẫn đến 0 với các giá trị $k > 6$.

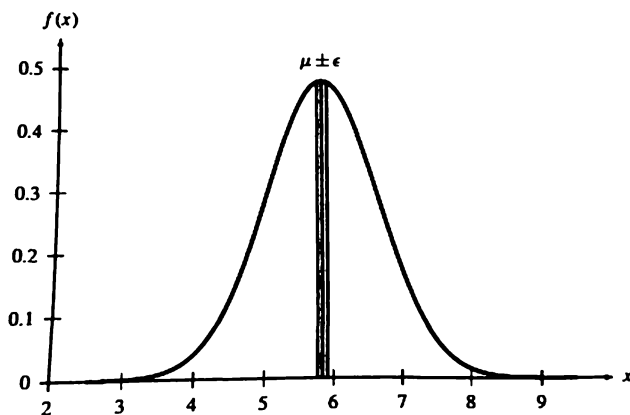
Hàm mật độ: khi xác suất nhỏ, không có nghĩa giá trị dữ liệu không tồn tại, mà do dữ liệu trong vùng quan sát thưa. Thay vì sử dụng hàm xác suất toàn thể, người ta xác định *hàm mật độ xác suất*, xác định xác suất để biến X lấy giá trị trong đoạn $[a, b]$:

$P(X \in [a, b]) = \int_a^b f(x) dx$. Hàm này không âm và tổng là 1, tức $\int_{-\infty}^{+\infty} f(x) dx = 1$.



Hình 2.35. Phân phối nhị thức: hàm khối xác suất, tức mật độ xác suất, với $m = 10, p = 0.087$

Người ta xét trong lân cận ε của x : $P(X \in [x - \varepsilon, x + \varepsilon]) = \int_{x-\varepsilon}^{x+\varepsilon} f(x)dx \approx 2\varepsilon \cdot f(x)$. Do vậy $f(x) \approx P(X \in [x - \varepsilon, x + \varepsilon]) / (2\varepsilon)$; có nghĩa mật độ xác suất tại x là tỉ lệ của xác suất và độ dài của đoạn, tức xác suất trên đơn vị độ dài. Khi xét lân cận hai điểm x_1, x_2 : $P(X \in [x_1 - \varepsilon, x_1 + \varepsilon]) / P(X \in [x_2 - \varepsilon, x_2 + \varepsilon]) \approx 2\varepsilon \cdot f(x_1) / (2\varepsilon \cdot f(x_2)) = f(x_1) / f(x_2)$. Điều này có nghĩa khi $f(x_1)$ lớn hơn $f(x_2)$ thì các giá trị X gần với x_1 xảy ra nhiều hơn so với gần x_2 .



Hình 2.36. Phân phối chuẩn: hàm mật độ xác suất, $\mu = 5.84, \sigma^2 = 0.681$

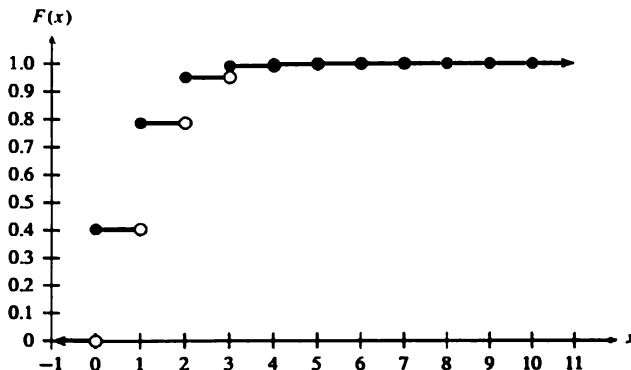
Về phân phối chuẩn, xét bảng dữ liệu. Giả sử các giá trị có hàm mật độ *chuẩn* hay *Gauss*, với $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$; trong đó μ là trung bình, σ^2 là biến thiên, phương sai, của các giá trị. Hình cho thấy hình chuông của phân phối chuẩn. Bảng dữ liệu cho biết $\mu = 5.84$; $\sigma^2 = 0.681$.

Trong khi $f(x=\mu) = f(5.84) = \frac{1}{\sqrt{2\pi\sigma^2}} = \exp\{0\} = 0.483$ người ta nhấn mạnh xác suất để quan sát $X = \mu$ là 0, tức $P(X = \mu) = 0$. Do vậy $P(X = x)$ không xác định theo $f(x)$, mà $P(X = x)$ như miền dưới đường cong đối với khoảng nhỏ $[x - \epsilon, x + \epsilon]$, tâm là x , với $\epsilon > 0$. Hình vẽ mô tả điều này, với miền gạch gạch có tâm là $\mu = 5.84$.

Theo công thức trên, $P(X = \mu) \approx 2\epsilon \cdot f(\mu) = 2\epsilon \cdot 0.483 = 0.967\epsilon$. Khi $\epsilon \rightarrow 0$, $P(X = \mu) \rightarrow 0$. Tuy nhiên, người ta thấy xác suất của các giá trị quan sát gần với $\mu = 5.84$ là 2.96 lần xác suất của các giá trị gần với $x = 7$, do $f(5.84)/f(7) = 0.483/0.18 = 2.69$.

Hàm phân phối tích lũy: đối với biến X , hàm phân phối tích lũy là xác suất quan sát các giá trị nhiều nhất là x : $F(x) = P(X \leq x) \forall -\infty < x < \infty$. Hàm F với X rời rạc là $F(x) = P(X \leq x) = \sum_{u \leq x} f(u)$, và với

X liên tục là $F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du$.



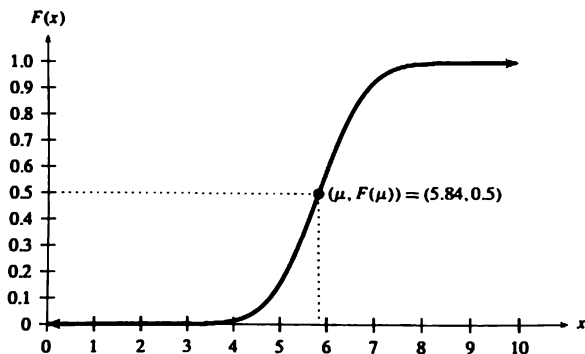
Hình 2.37. Hàm phân phối tích lũy đối với phân phối nhị thức

Về hàm phân phối tích lũy, hình cho thấy hàm phân phối tích lũy đối với phân phối nhị thức trong Hình 2.37. $F(x)$ có cùng giá trị như $F(k) \forall x \in [k, k+1]$, với $0 \leq k < m$; trong đó m là số các thử nghiệm, k là số các thử nghiệm thành công.

Hình 2.38 cho thấy hàm phân phối tích lũy đối với hàm mật độ chuẩn. Đối với biến liên tục, hàm phân phối tích lũy cũng liên tục, không giảm. Do phân phối chuẩn là đối xứng qua trung bình $F(\mu) = P(X < \mu) = 0.5$.

2.6.4.2. Các biến ngẫu nhiên hai biến

Thay vì xét mỗi thuộc tính như biến ngẫu nhiên, người ta phân tích cặp các thuộc tính X_1, X_2 , như *biến ngẫu nhiên hai biến*¹ $X = (X_1, X_2)$. $X: O \rightarrow R^2$ là hàm gán cặp cho mỗi kết quả trong không gian mẫu, cặp số thực, tức vectơ hai chiều $(x_1, x_2) \in R^2$. Như đối với trường hợp một biến, nếu kết quả là số thì mặc định đây là hàm đồng nhất.



Hình 2.38. Hàm phân phối tích lũy đối với phân phối chuẩn

Hàm xác suất liên kết toàn thể: hàm f cần thỏa mãn $f(x) = f(x_1, x_2) \geq 0 \forall x_1, x_2 \in (-\infty, \infty)$; và điều kiện tổng xác suất

$$\sum_x f(x) = \sum_{x_1} \sum_{x_2} f(x_1, x_2) = 1.$$

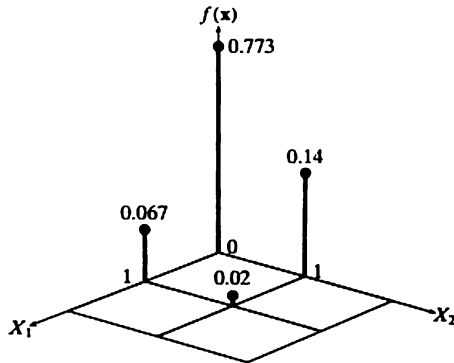
¹ Bivariate Random Variable: biến ngẫu nhiên hai biến.

Định nghĩa: Nếu X_1, X_2 là các biến ngẫu nhiên rời rạc, thì X có hàm xác suất liên kết toàn thể¹ $f(x) = f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(X = x)$.

Hàm mật độ xác suất liên kết: hàm f cần thỏa mãn $f(x) = f(x_1, x_2) \geq 0 \forall x_1, x_2 \in (-\infty, \infty)$; và điều kiện $\int_{R^2} f(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2$.

Tương tự như trường hợp một biến, $f(X_1) / f(X_2)$ cho phép người ta biết được xác suất một điểm gần với X_1 hay X_2 .

Định nghĩa: Nếu X_1, X_2 là các biến ngẫu nhiên rời rạc, thì X có hàm mật độ xác suất liên kết² là $P(X \in W) = \iint_{x \in W} f(x) dx = \iint_{(x_1, x_2)^T \in W} f(x_1, x_2) dx_1 dx_2$; trong đó W là tập con của không gian số thực hai chiều.



Hình 2.39. Hàm khối, tức mật độ, xác suất kết hợp, với X_1 là độ dài lá đài lớn, X_2 là độ rộng lá đài lớn

Về phân phối hai biến, xét độ rộng và độ dài đài hoa trong hình. Cho A là biến ngẫu nhiên Bernoulli ứng với độ dài cánh đài hoa, ít nhất 7 cm. Xác định biến B ngẫu nhiên Bernoulli khác, với chiều rộng cánh đài hoa, ít nhất 3.5 cm. Lấy $X = (A, B)$ là biến ngẫu nhiên hai biến, thì hàm xác suất liên kết toàn thể của X được ước lượng từ dữ liệu $f(0, 0) = P(A = 0, B = 0) = 116 / 150 = 0.773$; $f(0, 1) = P(A = 0, B = 1) = 21 / 150 = 0.140$; $f(1, 0) = P(A = 1, B = 0) = 10 / 150 =$

¹ Joint Probability Mass Function: hàm xác suất liên kết toàn thể.

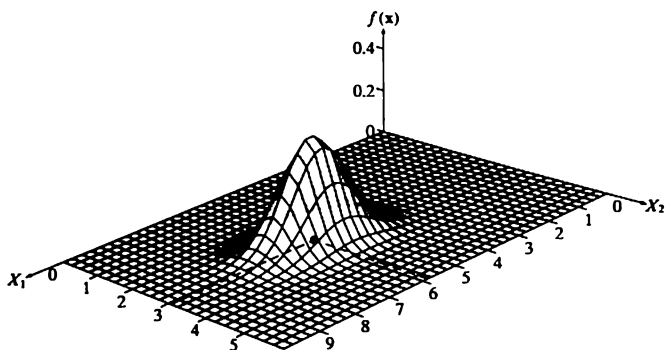
² Joint Probability Density Function: hàm mật độ xác suất liên kết.

0.067; $f(1, 1) = P(A = 1, B = 1) = 3/150 = 0.020$. Hình sau vẽ hàm xác suất này.

Xem các thuộc tính X_1, X_2 trong bộ dữ liệu. Như các biến ngẫu nhiên liên tục, có thể xác định biến $X = (X_1, X_2)$ ngẫu nhiên hai biến. Giả sử X theo *phân phối chuẩn hai biến*, hàm mật độ xác suất liên kết là $f(x|\mu, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left\{-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right\}$; trong đó

μ và Σ là các tham số của phân phối chuẩn hai biến, thể hiện vector trung bình hai chiều và ma trận hiệp phương sai. Kí hiệu $|\Sigma|$ chỉ định thức của Σ . Mật độ chuẩn hai biến như trong hình sau, với $\mu = (5.843, 3.054)^T$ và ma trận hiệp phương sai $\Sigma = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$.

Điều quan trọng là hàm $f(x)$ chỉ xác định mật độ xác suất tại x , và khác với $P(X = x)$. Thí dụ trên đã cho thấy $P(X = x) = 0$.



Hình 2.40. Mật độ chuẩn hai biến: $\mu = (5.843, 3.054)^T$, ứng với hình tròn đậm

Hàm phân phối tích lũy liên kết: hàm phân phối tích lũy liên kết đối với hai biến ngẫu nhiên X_1, X_2 được xác định là $F: F(x) = F(x_1, x_2) = P(X_1 \leq x_1 \text{ và } X_2 \leq x_2) = P(X \leq x), \forall x_1, x_2 \in (-\infty, \infty)$.

Độc lập thống kê: khi X_1, X_2 là độc lập, người ta có hàm phân phối tích lũy $F(x) = F(x_1, x_2) = F(x_1) \cdot F(x_2)$; và hàm mật độ $f(x) = f(x_1, x_2) = f(x_1) \cdot f(x_2)$.

Định nghĩa: Hai biến X_1, X_2 là độc lập thống kê nếu với mọi $W_1 \subset R$ và $W_2 \subset R$, $P(X_1 \in W_1 \text{ và } X_2 \in W_2) = P(X_1 \in W_1) \cdot P(X_2 \in W_2)$.

2.6.4.3. Biến ngẫu nhiên nhiều biến

Biến ngẫu nhiên nhiều biến d chiều $X = (X_1, \dots, X_d)^T$, hay *biến ngẫu nhiên vector*, là hàm gán kết quả vector số thực trong không gian mẫu, tức $X: \Omega \rightarrow \mathbb{R}^d$. Hạng của X được kí hiệu là vector $x = (x_1, \dots, x_d)^T$. Khi tất cả x_i là số, X gọi là hàm đồng nhất. Ngược lại, có giá trị phi số, X ánh xạ kết quả sang vector số trong hạng của nó.

Nếu X_i là rời rạc, thì X là rời rạc liên kết và hàm xác suất liên kết f: $f(x) = P(X = x)$; $f(x_1, \dots, x_d) = P(X_1 = x_1, \dots, X_d = x_d)$. Nếu X liên tục, thì X là liên tục liên kết, và hàm mật độ xác suất là $P(X \in W) = \int \dots \int_{X \in W} f(x) dx$; $P((X_1, \dots, X_d)^T \in W) = \int \dots \int_{(x_1, \dots, x_d)^T \in W} f(x_1, \dots, x_d) dx_1 \dots dx_d$ đối với miền $W \subseteq \mathbb{R}^d$ d chiều. Các luật về xác suất không âm và tổng xác suất bằng 1 vẫn đúng.

Hàm xác suất tích lũy của $X = (X_1, \dots, X_d)^T$ là $F(x) = P(X \leq x)$. $F(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d)$. Người ta nói X_i là các biến ngẫu nhiên độc lập nếu với mọi miền $W_i \subset \mathbb{R}^d$, $P(X_i \in W_i)$ và ... và $X_d \in W_d) = P(X_1 \in W_1) \dots P(X_d \in W_d)$. Khi các X_i độc lập, hàm phân phối tích lũy $F(x) = F(x_1, \dots, x_d) = F(x_1) \dots F(x_d)$; và hàm mật độ $f(x) = f(x_1, \dots, x_d) = f(x_1) \dots f(x_d)$.

2.6.4.4. Mẫu ngẫu nhiên và thống kê

Trong phân tích thống kê, người ta cần biết hàm xác suất toàn thể và hàm mật độ. Tuy nhiên không thể biết trước, nên người ta giả định phân phối nào đó, để tiếp tục công việc. Nhưng không thể xác định được các tham số trong phân phối đó. Do vậy cần ước lượng loại phân phối và các tham số từ dữ liệu.

Toàn bộ dữ liệu được coi như lực lượng, hay *dân cư*, của bộ dữ liệu. Việc phát hiện mô hình và tham số trên toàn bộ dữ liệu là (i) không khả thi; (ii) tốn kém. Do vậy cần sử dụng *mẫu*¹ đại diện.

Mẫu đơn biến: tất cả các S_i độc lập thống kê lẫn nhau, và theo cùng hàm xác suất toàn thể hay hàm mật độ. Hàm liên kết của

¹ Sample: mẫu.

chúng $f(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i)$; trong đó f_X là hàm xác suất toàn thể hay hàm mật độ.

Định nghĩa: Cho biến ngẫu nhiên X , mẫu ngẫu nhiên kích thước từ X là tập n biến ngẫu nhiên phân phối đồng nhất và độc lập¹ $S_1, \dots, S_n$.

Mẫu nhiều biến: để ước lượng tham số nhiều biến, n điểm dữ liệu x_i tạo nên mẫu ngẫu nhiên nhiều biến d chiều $X = (X_1, \dots, X_d)$. Tức là giả thiết x_i là độc lập và có phân phối đồng nhất, và phân phối liên kết của chúng là $f(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i)$; trong đó f_X hàm xác suất toàn thể hay hàm mật độ. Với giả thiết độc lập thì $f(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i) = \prod_{i=1}^n \prod_{j=1}^d f_{X_j}(x_{ij})$

Thống kê: có thể đánh giá tham số về lực lượng của dữ liệu nhờ xác định *thống kê* mẫu phù hợp, xem như hàm của mẫu. Cho $\{S_i\}$, $i = 1..m$ là các mẫu ngẫu nhiên kích thước m , rút ra từ biến ngẫu nhiên X .

Định nghĩa: Thống kê $\hat{\theta}$ là hàm $(S_1, \dots, S_m) \rightarrow R$.

Thống kê $\hat{\theta}$ là đánh giá của tham số lực lượng tương ứng. Bản thân $\hat{\theta}$ là biến ngẫu nhiên. Nếu sử dụng giá trị này để đánh giá tham số lực lượng, giá trị này được gọi là *ước lượng điểm*² của tham số, và thống kê này là *cái đánh giá* của tham số.

Về trung bình mẫu, xét thuộc tính độ dài đài hoa, X_1 , trong bộ dữ liệu hoa diên vĩ, trong bảng dữ liệu. Giả sử chưa biết giá trị trung bình của X_1 . Giả sử các giá trị quan sát $\{x_i\}$, $i = 1..n$ gồm mẫu ngẫu nhiên lấy từ X_1 . *Trung bình mẫu* là một thống kê, xác định theo $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$. Nhìn vào bảng dữ liệu, (

¹ IID: Independent and identically distributed variable: biến phân phối đồng nhất và độc lập.

² Point estimate: ước lượng điểm.

$\hat{\mu} = \frac{1}{150}(5.9 + \dots + 5.1) = 5.84$. Giá trị này là đánh giá điểm đối với tham số về lực lượng μ , giá trị trung bình của biến X_1 .

2.7. Quá trình chuẩn bị dữ liệu

Phân tiếp sau đề cập (i) khám phá dữ liệu như một quá trình; (ii) tác động của thể giới thực đến việc chuẩn bị dữ liệu; và (iii) chuẩn bị dữ liệu như một quá trình.

Dữ liệu cần thiết cho quá trình ra quyết định. Khám phá dữ liệu sẽ đảm bảo nguồn cơ bản cho quá trình ra quyết định, tức đảm bảo thông tin. Tuy nhiên không phải dữ liệu nào cũng có lợi cho khai phá dữ liệu. Giá trị của thông tin tùy thuộc vào quá trình lựa chọn dữ liệu phù hợp. Cũng vì vậy mà cần đến thời gian, tiền bạc, nguồn nhân lực, kỹ năng... để phát hiện thông tin phù hợp.

Người ta không cần tiếp thị cho quá trình chuẩn bị dữ liệu. Người ra quyết định hiểu rõ nhu cầu về thông tin, tại mọi mức ra quyết định khác nhau trong doanh nghiệp. Người ta sử dụng các công cụ để khám phá tri thức.

Những điều cần thiết đó cần được thực hiện theo quy trình, trong quá trình; đó là quá trình đảm bảo mọi người tham gia được huấn luyện, có kì vọng chung, đều cố gắng cao. Nó được gọi là quá trình khám phá dữ liệu.

2.7.1. Quá trình khám phá dữ liệu

Định nghĩa: Khám phá dữ liệu là quá trình kinh doanh nhiều giai đoạn thực hành, mà con người sử dụng phương pháp luận có cấu trúc để khám phá và đánh giá vấn đề phù hợp, xác định giải pháp và cài đặt chiến lược, sản xuất ra kết quả đo được.

Mỗi giai đoạn có mục tiêu và chức năng riêng. Vậy nên người ta cảm thấy cần biết cách xác định và điều cần làm tại mỗi giai đoạn. Nhìn nhận ở mức cao nhất, các giai đoạn của quá trình khám phá dữ liệu là:

1. Khám phá không gian vấn đề;
2. Khám phá không gian lời giải;
3. Xác định phương pháp thực hiện;
4. Khai phá dữ liệu, với các bước (i) chuẩn bị dữ liệu; (ii) khảo sát dữ liệu; (iii) mô hình hóa dữ liệu.

Các giai đoạn trên có vai trò, tức tầm quan trọng khác nhau. Tuy nhiên công sức, thời gian dành cho nhiệm vụ đó không hoàn toàn tương xứng.

			
Thời gian thực hiện (%)		Độ quan trọng để thành công (%)	
1. Khám phá bài toán	10	15	80 %
2. Khám phá lời giải	9	14	
3. Đặc tả thực hiện	1	51	
4. Khai phá dữ liệu			20 %
• Chuẩn bị dữ liệu	60	15	
• Khảo sát dữ liệu	15	3	
• Mô hình hóa dữ liệu	5	2	

Hình 2.41. Các giai đoạn của dự án khám phá dữ liệu, cho thấy độ quan trọng và thời lượng đối với mỗi giai đoạn

2.7.1.1. Giai đoạn khám phá không gian vấn đề

Đây là điểm căng thẳng cần làm, do dễ gây hiểu nhầm hay kỳ vọng sai cho việc khai phá dữ liệu. Khái niệm *khám phá dữ liệu* và *khai phá dữ liệu* không đồng nhất.

Việc khám phá dữ liệu hiệu quả cho phép tiếp cận chuyên môn để xác định bài toán kinh doanh và để hiểu dữ liệu dùng cho việc giải bài toán.

- *Xác định bài toán.* Vấn đề là cần xác định đúng bài toán, hay vấn đề cần giải. Việc đặt bài toán còn quan trọng hơn việc

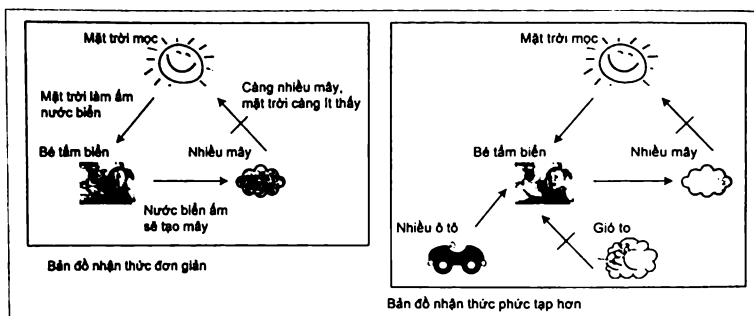
lập trình giải bài toán đó. Thuật toán tốt hay không còn tùy vào đầu bài có rõ ràng không. Tuy nhiên, dễ bị thuyết phục bởi thành công hiển hiện, mà người ta hay bỏ qua giai đoạn xác định bài toán. Chỉ khi biết rõ bài toán, người ta mới có mô hình phù hợp để sử dụng.

- *Xác định chính xác bài toán.* Bắt đầu với cách người ta xây dựng bài toán. Một chuyên gia hay nhóm chuyên gia có thể cảm thấy đã mô tả chính xác bài toán. Tuy nhiên, nói chung cần phải xem xét chi tiết, ứng với cấu trúc dữ liệu phù hợp. Có thể cần đến chuyên gia phản biện, để chi tiết hóa bài toán đã đặt ra. Công việc chi tiết hóa sẽ tạo nên nhiều bài toán con, yêu cầu nhiều công sức giải quyết. Thông thường việc xác định bài toán lớn lẫn bài toán con không thể dựa vào kĩ thuật tự động hóa, mà cần nhiều đến kĩ năng của phân tích viên.
- *Bản đồ nhận thức.* Đôi khi không muốn tượng được không gian bài toán. Nếu khó đi sâu vào cấu trúc của bài toán, hay khi gặp nhiều trường hợp xung đột, người ta có thể sử dụng tiếp cận khác để thể hiện không gian bài toán. Một phương pháp cho phép cấu trúc không gian bài toán là sử dụng công cụ *bản đồ nhận thức*.

Định nghĩa: *Bản đồ tư duy*¹ là phương pháp được đưa ra như là một phương tiện mạnh để tận dụng khả năng ghi nhận hình ảnh của bộ não.

Để khám phá không gian bài toán phức tạp, bản đồ nhận thức cho phép thể hiện hình ảnh vật lí của điều nhận thức được dưới dạng các đối tượng trong không gian bài toán, với các liên kết và tương tác giữa các biến đối tượng. Nó cũng cho biết vị trí các xung đột của cấu trúc bài toán.

¹ Mindmap: bản đồ tư duy.



Hình 2.42. Bản đồ nhận thức

Đối với các xung đột, điều quan trọng không phải là giải quyết hay loại bỏ xung đột, mà để hiểu chúng ra sao và chính xác xảy ra ở phần nào của bài toán. Thực tế, chúng có thể thể hiện diễn giải được kiểm chứng của các khung nhìn khác nhau của tình huống mà những chủ sở hữu bài toán sử dụng.

Xếp hạng quan trọng, theo cảm tính		
	Quay phim	3
	Nghe nhạc	1
	Nghe đài	2
	Xem phim	4
	Du lịch	5

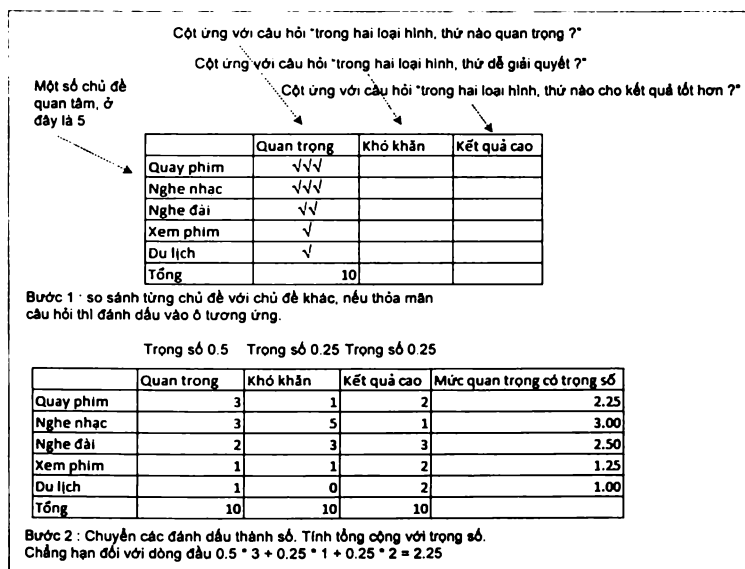
Hình 2.43. Xếp hạng trực quan, theo cảm tính

- *Lời giải cảm tính.* Khi bài toán đã rõ, đã được khám phá, điều quan trọng là sử dụng kĩ thuật *giải cảm tính*¹. Tiếp cận này phủ dãi rộng của lĩnh vực và kĩ thuật, với mục đích là đảm bảo hình ảnh thần kinh về bài toán trong đầu người

¹ Ambiguity resolution: giải mơ hồ.

chủ bài toán, tức hình ảnh thần kinh với nhiều giả thiết liên quan, được liên lạc với nhau rõ ràng, được người giải bài toán hiểu, tức làm sáng tỏ các giả thiết liên quan với nhau. Việc giải cảm tính dùng để đảm bảo nơi có các diễn giải khác, nơi giải thích được các giả thiết;

- *Xếp hạng theo cặp và xây dựng ma trận bài toán.* Khám phá không gian bài toán, tùy theo phạm vi dự án, đưa đến nhiều bài toán khả thi. Vấn đề là bài toán nào đáng giải, với thời gian và tài nguyên hạn chế.



Hình 2.44. Xếp hạng theo cặp

Trong lĩnh vực của lí thuyết quyết định và kinh tế lượng, người ta xác định được một cách tin cậy các bài toán hiệu quả và thích hợp, tức bài toán *xếp hạng theo cặp*¹. Việc tạo xếp hạng theo cặp là kĩ thuật mạnh để giảm các lựa chọn phải so sánh. Nó cho kết quả khác so với xếp hạng trực quan. So sánh theo cặp là quá

¹ Pairwise ranking: xếp hạng theo cặp, do L. L. Thurstone đề xuất năm 1927.

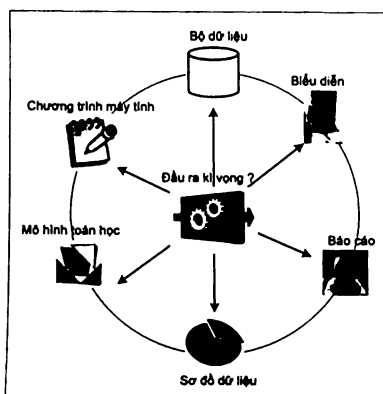
trình so sánh các thực thể theo cặp, để kiểm chứng nên chọn thực thể nào, hay có giá trị định lượng tốt hơn, hay so sánh hai thực thể khớp nhau không. Phương pháp so sánh theo cặp được dùng trong nghiên cứu khoa học, hệ thống bầu chọn, lựa chọn xã hội, hệ thống trí tuệ nhân tạo đa tác nhân.

Ma trận các lựa chọn được sử dụng để tiện so sánh. Việc tạo ma trận thế giới thực là phức tạp, đặc biệt đối với bài toán có nhiều chủ sở hữu. Việc so sánh toàn bộ sẽ không khả thi, do sự phức tạp tính toán.

2.7.1.2. Giai đoạn khám phá không gian lời giải

Sau khi xác định được bài toán cần giải, và xếp hạng chúng phù hợp, người phân tích cần giải chúng. Trước khi thu về lời giải, hay nghiệm, của bài toán, cũng nên biết nó ra sao.

Đầu ra điển hình từ khám phá dữ liệu đơn giản gồm (i) báo cáo; (ii) đồ thị; (iii) mã chương trình; (iv) danh sách các bản ghi; (v) công thức đại số... Nhiều bài toán có cùng chung lời giải.



Hình 2.45. Cần biết đầu ra phù hợp ra sao với không gian lời giải

Không thể khẳng định được trong giai đoạn khám phá không gian lời giải, người ta thu được nghiệm chính xác đối với thế giới thực. Đặc tả ở giai đoạn này là cần thiết cho quá trình khám phá

dữ liệu, chứ không cho quá trình khai phá dữ liệu. Việc khai phá dữ liệu cho ra lượng hạn chế kết quả, nhưng mỗi cái cần khớp với nhu cầu chung.

Trong thực tế, đôi khi người ta thấy vấn đề hợp lý và xác định mô hình cho nó, hoặc lấy một mô hình rồi cải tiến để đáp ứng nhu cầu của tổ chức. Điều này cho thấy khó xác định được phương pháp luận để xác định rõ lời giải cần tìm cho chính xác.

Cả khám phá bài toán lẫn khám phá lời giải đều cần đến tiếp cận giải mơ hồ, tức xấp xỉ. Đây là kỹ thuật dùng để thử nghiệm về điều liên quan và xác định cái cần đến. Kỹ thuật mơ hồ tìm kiếm phần chưa hiểu trong truyền thông, phát hiện các giả thiết nền, và đảm bảo hiểu các ý chính của bài toán. Loại bỏ sự mơ hồ trong khám phá dữ liệu thế giới thực cũng phức tạp.

2.7.1.3. Giai đoạn xác định phương pháp cài đặt

Khi đã xác định các bài toán và xếp hạng chúng, đã xác định hình dạng lời giải, các kì vọng và đặc tả đã khớp và các giả thiết ẩn đã lộ ra, bắt đầu chuyển qua giai đoạn *cài đặt phương pháp*.

Ở giai đoạn này sẽ không còn phải khám phá điều mới nữa, mà cần để cập (i) tăng lợi ích; (ii) cải thiện hiệu năng; (iii) tăng mức thỏa mãn của khách hàng; (iv) giảm lãng phí; (v) giảm lỗi; (vi) đáp ứng đích kinh doanh khác. Điều này đòi hỏi đầu tư cho dự án khám phá dữ liệu, để thay đổi hành vi hệ thống.

Đặc tả cài đặt, hay thực hiện, là bước cuối cùng liên quan đến cách các lời giải khác nhau đối với bài toán đã chọn sẽ được ứng dụng trong thực tế. Việc này chi tiết hóa khuôn mẫu cuối cùng cho dự án. Đặc tả cần được xác định hoàn toàn thực tế, tức (i) bài toán nhằm vào điều gì; (ii) có khuôn mẫu nào; (iii) phân phối giá trị ra sao; (iv) người sẽ sử dụng chúng; (v) sản xuất chúng ra sao; (vi) hạn chế và kì vọng; (vii) chu kì sống...

Ở giai đoạn này, cần xác định rõ (i) *người sở hữu bài toán*, tức người có kinh nghiệm về bài toán hiện tại; (ii) *người giữ bài toán*,

tức người điều khiển tài nguyên cho phép thực hiện lời giải bài toán. Tài nguyên có các dạng (i) tiền; (ii) nhân lực; (iii) thời gian; (iv) chính sách hợp tác... Để hiệu quả, lời giải đưa ra cần hiệu quả về giá và đáp ứng người nắm bài toán.

2.7.1.4. Giai đoạn khai phá dữ liệu

Khai phá dữ liệu là quá trình nhiều bước nhỏ (i) chuẩn bị; (ii) khảo sát, xác định lại dữ liệu; (iii) mô hình hóa.

Nội dung về các bước nhỏ đã chi tiết hóa trong các chương khác, nên không trình bày lại. Ở đây chỉ nêu các ý chính:

- *Chuẩn bị dữ liệu cho mô hình hóa.* Dữ liệu là cần thiết; tuy nhiên cần loại bỏ những dữ liệu thừa, rác. Tuy có phương pháp luận về thuật toán và mô hình, nhưng không có phương pháp luận hay kinh nghiệm về kỹ thuật chuẩn bị dữ liệu. Dữ liệu chuẩn bị tốt là dữ liệu thực hành được với mô hình trong thế giới thực. Việc chuẩn bị dữ liệu dừng khi bộ dữ liệu đủ cho mô hình;
- *Môi trường thông tin đã chuẩn bị PIE¹.* Đích thứ hai của chuẩn bị dữ liệu là sinh ra môi trường thông tin đã chuẩn bị PIE. Nó là chương trình máy tính động, làm chương trình bao đối với công cụ mô hình hóa, để tránh dữ liệu hư hại hay biến dạng;
- *Khảo sát dữ liệu.* Liên quan đến khảo sát dữ liệu có ba câu hỏi chính (i) bộ dữ liệu là gì; (ii) có thể thu được kết quả không; (iii) đâu là vùng hạn chế, nguy hiểm. Các câu hỏi này cũng tương tự như câu hỏi đặt ra đối với quá trình mô hình hóa, nhưng với ý nghĩa khác;
- *Mô hình hóa dữ liệu.* Nếu chú trọng vào công cụ mô hình hóa khai phá dữ liệu, xem nó như tiếp cận đầu tiên để giải bài toán, thường sẽ dẫn đến các bài toán được hình thức

¹ PIE: prepared Information Environment: môi trường thông tin đã chuẩn bị.

hóa không thích hợp. Do vậy nên đặt ra câu hỏi (i) cách người ta có thể làm...; (ii) lí do như vậy....

Như vậy, sau một số giai đoạn, có thể thấy được khám phá¹ như là khai phá dữ liệu và mô hình hóa. Quá trình khám phá dữ liệu không nhấn mạnh bất kì giai đoạn nhỏ nào, mà chúng đều có ý nghĩa quan trọng. Cần xác định các bài toán, rồi xác định các lời giải tiềm năng. Các giai đoạn đều liên hệ với nhau. Giai đoạn mô hình hóa, gồm các công cụ, liên quan nhiều đến dạng dữ liệu đã chuẩn bị.

2.7.2. Khai phá dữ liệu, mô hình hóa và các công cụ mô hình hóa

Một đích chính của chuẩn bị dữ liệu là phục vụ cho việc phát hiện mô hình. Chương trước đã đề cập mô hình, ứng với toàn thể dữ liệu, và khuôn mẫu, ứng với phần nhỏ của bộ dữ liệu.

2.7.2.1. Quy tắc quan trọng

Việc khám phá dữ liệu giúp xây dựng khung để khai phá dữ liệu, sao cho các công cụ phù hợp dùng cho dữ liệu thích hợp, đã được chuẩn bị, nhằm giải vấn đề kinh doanh và giải pháp cần phân phối. Liên quan đến khung này, người ta đề xuất mười quy tắc vàng.

1. Lựa chọn bài toán được xác định rõ ràng, cho phép thu được lợi ích hữu hình;
2. Xác định lời giải theo yêu cầu;
3. Xác định cách để lời giải sẽ được sử dụng;
4. Hiểu về bài toán và bộ dữ liệu, tức lĩnh vực quan tâm;
5. Cho phép bài toán dùng mô hình, tức lựa chọn công cụ, chuẩn bị dữ liệu...;
6. Định ra các giả thiết;
7. Làm tinh mô hình nhiều lần;

¹ Exploration: khám phá; Mining: khai phá (dữ liệu).

8. Chỉnh mô hình về dạng đơn giản nhất, nhưng không tầm thường;
9. Xác định điều bất ổn trong mô hình, tức xác định miền căng mà tại đó thay đổi một chút đầu vào vẫn khiến thay đổi lớn ở đầu ra;
10. Xác định tính không chắc chắn của mô hình, tức xác định miền căng và phạm vi khiến mô hình sinh ra dự đoán ít tin cậy.

Liên quan đến đầu ra của quá trình, biểu thức cuối cùng của các mối quan hệ có thể được thể hiện dưới nhiều dạng, thông dụng là (i) đồ thị; (ii) đẳng thức toán học; (iii) chương trình máy tính. *Mô hình bị động* diễn tả mối quan hệ phát hiện thấy trong bộ dữ liệu; nó có dạng đồ thị, biểu đồ. *Mô hình chủ động* có cùng đầu vào và cho ra dự đoán kì vọng.

2.7.2.2. Về các công cụ mô hình hóa

Ở đây không chi tiết về các công cụ mô hình hóa trong khai phá dữ liệu. Nội dung này đã có trong các chương trước. Các công cụ mô hình hóa liên quan đến các chương trình máy tính, thực hiện việc diễn giải và xử lí bộ dữ liệu. Các công cụ có thể là (i) tiện ích cho việc mô hình hóa; (ii) xử lí dữ liệu trong cơ sở dữ liệu và trong kho dữ liệu; (iii) phân tích thống kê. Chúng đều sử dụng tri thức lĩnh vực.

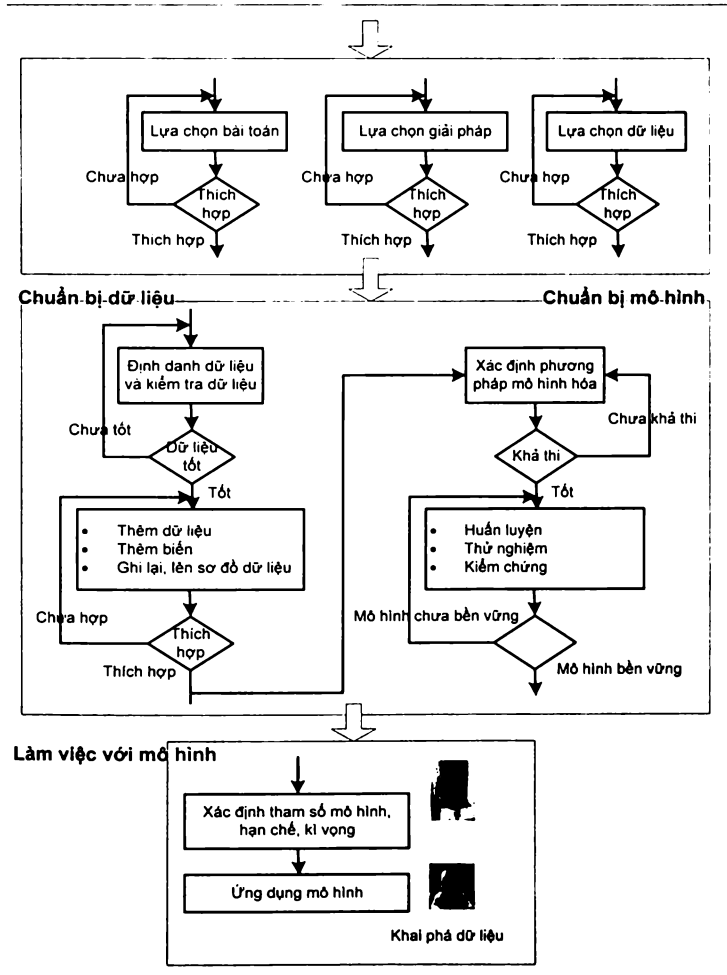
Việc phân tích thống kê hướng về việc kiểm tra và kiểm định các giả thuyết. Phân tích thống kê dùng để phân tích dữ liệu vừa khám phá; nó phù hợp với suy nghĩ của con người, thể hiện kết quả dưới dạng đồ họa và lượng hóa mối quan hệ vốn có trong dữ liệu, cho phép tìm kiếm theo mẫu.

2.7.2.3. Các dạng mô hình

Cần đến các mô hình khi đã định được (i) bài toán; (ii) tập giải pháp kì vọng; (iii) chiến lược thực hiện; (iv) chuẩn bị dữ liệu; (v) khảo sát dữ liệu; (vi) lựa chọn thuật toán. Người ít kinh

liên quan về mô hình hóa nhằm tường thuật quá trình mô hình hóa đến tính. Quá trình tuyến tính thường là (i) chọn công cụ; (ii) lựa chọn dữ liệu; (iii) tạo mô hình; (iv) áp dụng mô hình với dữ liệu; (v) đánh giá kết quả.

Cần xác định được việc xây dựng mô hình là quá trình liên tục, có vòng phản hồi lặp lại, có tương tác giữa nhiều thành phần.



Hình 2.46. Các nhiệm vụ xây dựng mô hình

2.7.2.4. Mô hình chủ động và mô hình bị động

Hai mô hình vừa được đề cập: (i) mô hình chủ động sẽ đáp ứng theo cách nào đó; (ii) mô hình bị động không có phản ứng.

Định nghĩa: Mô hình bị động là mô hình trả lời các câu hỏi và thể hiện mối quan hệ với biểu đồ, đồ thị, văn bản, công thức toán học.

Mô hình bị động không lí giải về mối quan hệ. Nó hoạt động theo một cách đã xác định.

Định nghĩa: Mô hình chủ động là mô hình thực hiện nhiều hoạt động, chủ động yêu cầu dữ liệu đầu vào.

Phân biệt hai loại mô hình cũng phức tạp đối với ứng dụng và người dùng. Người ta tách biệt chúng theo dữ liệu dùng để mô hình hóa. Khi chuẩn bị dữ liệu, khác nhau giữa hai mô hình không nhiều, nếu chúng tác động ít đến cách chuẩn bị dữ liệu để mô hình hóa.

2.7.2.5. Mô hình giải thích và mô hình dự đoán

Người ta thấy có (i) mô hình được tạo để giải thích các luồng dữ liệu, đó là *mô hình giải thích*; (ii) mô hình được thiết kế để dự đoán, phân lớp và diễn giải dữ liệu, đó là *mô hình dự đoán*.

Cả hai mô hình này đều có dạng (i) chủ động, hay (ii) bị động. Chẳng hạn trong tự động hóa công nghiệp, đầu ra của quá trình mô hình hóa là *mô hình dự đoán bị động*; ngược lại với bài toán kinh doanh, không thể tự động được, mà chỉ *tự động lấy ý kiến chuyên gia*. Tiếp cận mô hình hóa với bài toán kinh doanh yêu cầu theo dõi và mô hình hóa điều chuyên gia đã làm, chứ không yêu cầu chuyên gia tạo ra mô hình.

2.7.2.6. Mô hình học liên tục và mô hình tĩnh

- *Mô hình tĩnh.* Dữ liệu lịch sử trong mô hình tĩnh là các dữ liệu dùng để xây dựng mô hình, và chúng không được cập nhật. Người ít kinh nghiệm về mô hình hóa thường chỉ

thấy các mô hình tĩnh. Mô hình tĩnh cũng được sử dụng cho nhiều bài toán thực tế.

Định nghĩa: Mô hình tĩnh là mô hình khám phá mối quan hệ hay trả lời câu hỏi từ các dữ liệu lịch sử.

- *Mô hình học liên tục.* Loại mô hình này thể hiện những thứ tương đối, có điều khiển và làm việc với môi trường động.

Định nghĩa: Mô hình học liên tục là hệ thống sử dụng mô hình tự trị, với nhiều đầu môi bên trong, tự hoàn thiện trong suốt quá trình làm việc.

Chẳng hạn con người là thí dụ về nơi diễn ra quá trình học liên tục. Trong hệ thống học liên tục nhân tạo, các đầu môi ban đầu do bên ngoài xác định; quá trình học sẽ hoàn thiện các đầu môi của hệ thống.

Liên quan đến chuẩn bị dữ liệu, các hoạt động đầu tiên làm việc với mô hình tĩnh. Sau đó người ta sử dụng mô hình học liên tục. Khi đó môi trường thông tin đã chuẩn bị (PIE) cho phép việc chuẩn bị dữ liệu tự động, liên tục và là phần quan trọng trong quá trình.

2.8. Bản chất của thế giới và tác động của nó với việc chuẩn bị dữ liệu

Dữ liệu được khám phá để phát hiện tri thức trên dữ liệu, tức tri thức về thế giới. Như vậy phải định lượng được tri thức qua các độ đo và phân biệt (i) dữ liệu; (ii) thông tin và (iii) tri thức.

2.8.1. Đo về thế giới

Thế giới phức tạp; người ta đơn giản nó để hiểu và sống chung với nó. Việc đơn giản hóa cho phép thể hiện thế giới, đo các dữ liệu, tức các biểu hiện một số khía cạnh của thế giới.

2.8.1.1. Các đối tượng

Thế giới gồm các đối tượng. Trong công nghệ thông tin, người ta xác định nhiều thuật ngữ; tuy nhiên có một số thuật ngữ khiến

người ta ngại định nghĩa rõ ràng. Chẳng hạn đối tượng được xác định như thực thể, cái được quan tâm, nhưng không xác định rõ được nó ra sao. Tiếp cận hướng đối tượng lấy đối tượng là khái niệm cơ sở. Vậy nên xem tất cả những thứ trong thế giới thực, xung quanh con người, đều là các đối tượng. Đối với khám phá dữ liệu, đối tượng là dạng cơ bản cần khám phá. Người khai phá dữ liệu xem đối tượng theo góc nhìn thực nghiệm mà không xác định chính xác đối tượng, và chỉ xem đối tượng là tập các *đặc trưng* có thể đo được.

Đối tượng không cần thiết ở dạng vật lí. Chẳng hạn giá sinh hoạt là đối tượng không vật lí hóa được. Nó cũng có định nghĩa và các đặc trưng. Các đặc trưng của giá sinh hoạt là đo được, chẳng hạn như tiền thu nhập, tỉ lệ thay đổi theo tháng...

Các đối tượng tương tác với nhau. Các đặc trưng của đối tượng ở dạng dữ liệu sẽ phản ánh hệ thống về thế giới. Nếu phản ánh này là chính xác, tự các đặc trưng thể hiện hệ thống.

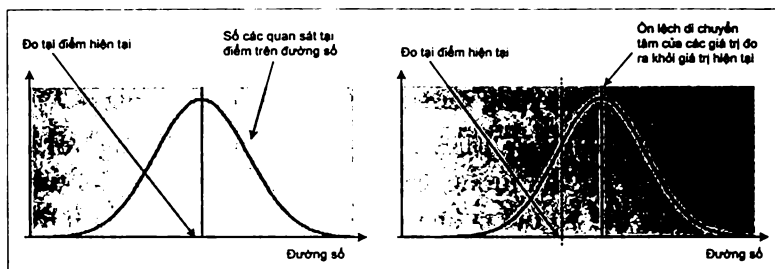
2.8.1.2. Độ đo

Các đối tượng được thể hiện qua các độ đo đặc trưng. Một nhóm các độ đo đặc trưng sẽ xác định tính chất của đối tượng, đại diện cho một thể hiện của đối tượng. Số đo đặc trưng cần được kiểm chứng, tức phải mang một số giá trị đã định. Đặc trưng kiểm định mang yếu tố thời gian, tức có dấu thời gian.

Người ta giả sử rằng các số đo về đối tượng đã được kiểm chứng, tại hoàn cảnh xác định.

2.8.1.3. Sai số độ đo

Tương tự đơn giản việc đo bằng thước. Lỗi sẽ gặp (i) mức chuẩn của thang đo trên cái thước; (ii) đối tượng được đo bằng thước ở trạng thái không bình thường; (iii) tuy đo bằng thước, nhưng nhìn số đo sai, dẫn đến kết quả sai...



Hình 2.47. Ổn thiên lệch khiến hầu hết số đo cụm lại quanh điểm được thực sự đo

2.8.1.4. Đưa số đo vào thế giới thực

Đôi khi số đo được mô tả gồm (i) giá trị hoàn hảo tuyệt đối thực sự; (ii) thêm vào độ xoắn, vặn, tức biến dạng. Độ vặn thường xem như sai số đo. Tuy nhiên, độ vặn thực ra là thuộc về số đo, chứ không có lỗi hay sai sót ở đây.

Ngược với viễn cảnh với số đo hoàn hảo với sai số, tình huống thực tế hơn là khớp độ đo vào thế giới thực, cộng thêm số ước lượng ứng với hoàn cảnh thực.

2.8.2. Các loại số đo

Xuất phát từ sự khác nhau giữa số đo 1.78 m với số đo màu xanh, người ta thấy nên xác định các loại độ đo. Tất cả các độ đo đều có đặc trưng chung: chúng dựa vào thang đo. Việc đo nói chung là khớp thế giới theo cách thể hiện qua số đo.

Liên quan đến biến ngẫu nhiên, chúng có hai thành phần (i) thang đo để tự đo được; (ii) thành phần gộp với nhiều thành phần. Lớp các biến có thể được chỉ thị qua vị trí của điểm đơn, tức giá trị, trên thang đo cụ thể, được gọi là các *biến vô hướng*. Các loại biến cần nhiều giá trị mới xác định được, được gọi là *biến vector*.

2.8.2.1. Độ đo vô hướng

Người ta dùng nội dung thông tin để phân loại các kiểu độ đo vô hướng. Một số loại là:

1. *Độ đo vô hướng theo tên.* Các giá trị đo là các tên, tức tên gọi của đối tượng. Nó được dùng để định danh đối tượng. Cách đo này cũng xem như đánh nhãn đối tượng;
2. *Độ đo vô hướng phạm trù.* Số đo phạm trù ứng với một nhóm các đối tượng, chứ không ứng với một đối tượng, thực thể. Người ta không biết về kiểu dữ liệu, kích thước nhóm; vậy nên độ đo này cũng như đánh nhãn cho nhóm đối tượng;
3. *Độ đo vô hướng có thứ tự.* Với độ đo này, các đối tượng được sắp xếp, cho biết vị trí đối tượng trong danh sách đối tượng. Độ đo này còn được gọi là *độ đo thứ tự*. Độ đo thứ tự cho biết nhiều thông tin hơn độ đo theo tên, hay độ đo phạm trù;
4. *Độ đo khoảng vô hướng.* Độ đo thứ tự cho biết sự phân hạng giữa các đối tượng. Còn thông tin khác, liên quan đến kích thước, cho biết sự khác nhau giữa các đối tượng theo kích thước. Đó là thang đo *khoảng*. Do độ đo này liên quan đến sự phân biệt, về sự thay đổi, hay số định lượng của thuộc tính đối tượng, người ta gọi nó là *độ đo định lượng*;
5. *Độ đo tỉ lệ vô hướng.* Độ đo này mang nhiều thông tin nhất. Nó bắt đầu đếm từ điểm 0. Độ đo tỉ lệ cũng là độ đo định lượng, với hai dạng (i) có thang đo đo được, có gán tên; (ii) không có thang đo gán tên. Đặc trưng của mỗi dạng có ý nghĩa khi chuẩn bị dữ liệu.

2.8.2.2. *Độ đo có hướng*

Với độ đo có hướng, cần hai thành phần (i) giá trị đo; (ii) thang đo. Chẳng hạn tốc độ xe đi kèm với thang km/giờ. Độ đo có hướng cần nhiều thành phần để có thêm thông tin. Chẳng hạn độ đo tốc độ có thể được vẽ theo đồ thị hai chiều, dưới dạng điểm xác định bởi độ đo tốc độ và trục khác là hướng đi.

Khái niệm về không gian trạng thái quan tâm đến giá trị đo của nhiều đặc trưng, như vector nhiều chiều.

2.8.3. Mở rộng đối với thuộc tính của biến

Các độ đo sử dụng các dạng khác nhau về thang đo. Tập các giá trị đo của đặc trưng cụ thể được nhóm thành các biến. Do các giá trị đo gộp lại, có thể nhìn nhận các mẫu theo cách các giá trị thay đổi trong đặc trưng kiểm định, hay sự thay đổi của biến khác. Sau đây là khái niệm về *mở rộng*¹ đối với thuộc tính của các biến.

2.8.3.1. Mở rộng về định tính, định lượng

Mở rộng này giữ nội dung thông tin thấp đến cao của các dạng khác nhau của các biến vô hướng. Việc mô tả biến theo định tính hay định lượng là không nhất thiết để mô tả nội dung thông tin. Các biến định danh, tức biến tên, là định tính.

2.8.3.2. Mở rộng về rời rạc, liên tục

Cần phân biệt các loại biến. Thực ra, phân biệt rời rạc với liên tục là sự mở rộng. Người ta dễ hình dung nếu xác định đối tượng theo các điểm trên chiều mở rộng đó. Các biến rời rạc được xem như có tập hạn chế các giá trị.

- *Biến giá trị đơn*. Biến này chỉ nhận giá trị đơn, nhưng thay đổi. Nếu biến không thay đổi giá trị, thì nó không có ý nghĩa đối với xây dựng mô hình;
- *Biến hai giá trị*. Biến này còn được gọi là *biến phân đôi*²;
- *Giá trị thiếu và giá trị rỗng*. Vấn đề cần đề cập trong chuẩn bị dữ liệu là thiết dữ liệu. Có hai giá trị liên quan (i) *thiếu*; (ii) *rỗng*³.

Định nghĩa: Giá trị thiếu đối với một biến là giá trị không tham gia vào bộ dữ liệu, nhưng tồn tại trong thực tế, có số đo.

¹ Continuum: liên tục, mở rộng.

² Dichotomy: phân đôi.

³ Missing: thiếu; empty: rỗng.

Định nghĩa: Giá trị rỗng trong một biến là giá trị mà không có trong thế giới thực, để cung cấp. Các giá trị thiếu và rỗng được gọi là trống¹. Có hai dạng giá trị trống (i) ứng với giá trị thiếu; (ii) ứng với giá trị rỗng. Nó không phải số đo;

- *Biến nhị phân.* Đó là dạng biến phân đôi. Giá trị biến nhị phân là 0, 1. Các giá trị này còn được xem là điều kiện *đúng, sai*,
- *Biến liên tục.* Biến liên tục bị hạn chế bởi giá trị cực đại và cực tiểu, ít nhất là theo lý thuyết, tức nó nhận giá trị trong khoảng. Chẳng hạn nhiệt độ là biến liên tục.

Bảng 2.11. Các dạng thang đo

Kiểu vô hướng	Độ đo	Giá trị đo	Ghi chú
Định danh, tức theo tên	Tên	Mờ	Phân biệt đối tượng này với đối tượng khác.
Phạm trù	Giống		Có thể chọn phạm trù khác.
Phạm trù phân đôi	Giới tính	Nam, nữ	
Phạm trù nhị phân	Đúng, sai	1, 0	
Phạm trù, thiếu	Màu da	Giá trị có trong thế giới thực.	
Phạm trù, rỗng	Số xe	Chưa có dữ liệu trong thực tế.	
Thứ tự	Độ dài	Dài, ngắn	Độ dài so sánh cho phép xếp hạng.

Liên quan đến các biến là khái niệm về dữ liệu, thông tin, và giá trị số đo của chúng. Người ta khai phá dữ liệu, từ đó khám phá thông tin, phát hiện tri thức. Cả dữ liệu, thông tin, tri thức đều cần cho quá trình phân tích thống kê và khai phá dữ liệu. Vấn đề chuyển dạng thể hiện cũng quan trọng; người ta chuyển dạng dữ liệu này sang dữ liệu khác, và chuyển hóa sang dạng khác là thông tin, tri thức.

2.8.4. Xây dựng các dạng biểu diễn của dữ liệu, tiện cho khám phá

Để sử dụng biến cho việc khai phá dữ liệu, cần đề cập dạng của dữ liệu. *Dữ liệu* là giá trị thể hiện đơn của một biến. *Số đo* có nghĩa cả dữ liệu, lẫn giá trị của đặc trưng, của biến, theo điều kiện kiểm chứng.

Kiểm chứng hiện tượng đối với dữ liệu là đặc trưng nội tại của dữ liệu, là cách để cấu trúc hóa dữ liệu. Tất cả những giả thiết nên, về nội tại, và các lựa chọn sẽ tạo nên bộ dữ liệu.

2.8.4.1. Biểu diễn dữ liệu

Dữ liệu được thể hiện theo nhiều dạng (i) bảng tính; (ii) dạng đồ họa như biểu đồ, đồ thị; (iii) ma trận. Trong máy tính, dữ liệu được chứa trong các tệp. Tổ chức tệp đa dạng (i) tuần tự; (ii) có chỉ mục; (iii) ngẫu nhiên.

Thể hiện dữ liệu có ý nghĩa đối với thuật toán, tức thuật toán xây dựng trên cấu trúc dữ liệu cụ thể.

2.8.4.2. Xây dựng dữ liệu qua các biến

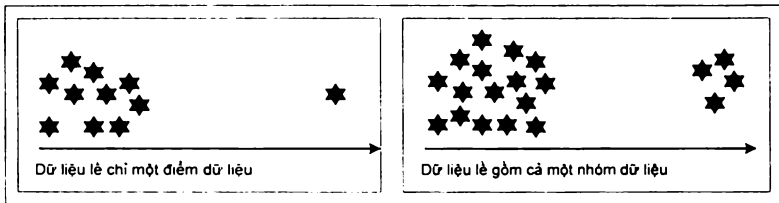
Việc thể hiện dữ liệu đề cập (i) dữ liệu; (ii) bộ dữ liệu. *Dữ liệu* nhằm vào các biến, như thực thể riêng; tương tác hay mối quan hệ giữa các biến là phụ. *Bộ dữ liệu* nhằm vào không chỉ các biến, mà tương tác và mối quan hệ giữa chúng. Việc khai phá dữ liệu tạo mô hình và chỉ làm việc với các bộ dữ liệu. Vậy nên việc chuẩn bị dữ liệu liên quan đến các biến riêng và cả bộ dữ liệu.

Các biến có đặc trưng theo một số cách, liên quan đến hành động để chuẩn bị dữ liệu.

1. *Biến như các đối tượng.* Điều quan trọng là các số đo với biến cần được kiểm chứng. Các suy diễn trên đặc trưng của biến cũng như làm việc với biến hiện tại. Đối với biến, dữ liệu đủ còn mang ý về các đặc trưng của biến cần được kiểm soát;
2. *Loại bỏ biến.* Khi đo đặc trưng, cần đếm số các giá trị thể hiện. Trong mẫu bất kì của các giá trị, chỉ có hữu hạn giá trị khác nhau, tức kích thước mẫu. Các biến (i) chỉ gồm giá trị thiếu; hay (ii) chỉ có một giá trị, cần bị loại khỏi quá trình mô hình hóa;
3. *Tính thừa.* Khi lẽ ra biến có nhiều giá trị, nhưng thiếu, nên dữ liệu thừa; người ta loại bỏ biến. Tuy nhiên với biến có bản chất là thừa dữ liệu, thì người ta có biện pháp xử lí các bộ dữ liệu thừa, để tiện cho (i) lưu trữ; (ii) thuật toán trên dữ liệu thừa;
4. *Tính đơn điệu.* Biến đơn điệu là biến có xu thế tăng/giảm như nhau cho đến biên giá trị. Tính đơn điệu cũng xuất hiện trong mối quan hệ giữa các biến.
5. *Việc tăng chiều.* Bộ dữ liệu nhiều chiều được chuyển về ít chiều, để giảm độ phức tạp, nhưng vẫn phải giữ được đặc trưng chính;
6. *Dữ liệu lẻ.* Dữ liệu lẻ¹ là các xuất hiện các giá trị với tần suất thấp theo một biến đơn, mà cách xa cụm giá trị của biến này. Đương nhiên, chúng ảnh hưởng đến độ chính xác của mô hình tạo nên từ dữ liệu, tức lệch khỏi đồng dữ liệu chính;
7. *Đánh số giá trị phạm trù.* Xử lí đúng các giá trị phạm trù là một trong nhiệm vụ quan trọng trong chuẩn bị dữ liệu. Một số trường hợp, người ta chuyển giá trị phạm trù sang

¹ Outlier: lẻ.

dạng số, theo cách không quan tâm đến các giá trị phạm trù không mang ý nghĩa. Chẳng hạn đối với giá trị độ cao là cao, thấp, trung bình, chuyển thành 3, 1, 2.

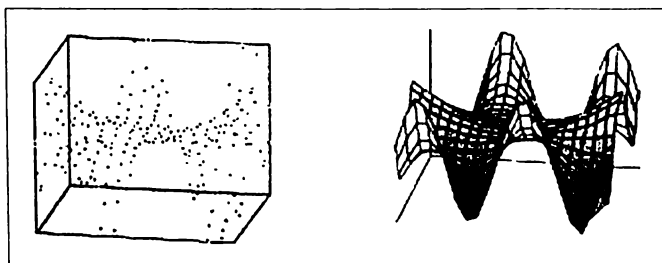


Hình 2.48. Dữ liệu lẻ

2.8.4.3. Xây dựng bộ dữ liệu để khai phá

Với bộ dữ liệu, cần để ý đến mối quan hệ giữa các biến. Đó là cách tự nhiên về quan hệ lẫn nhau giữa các biến. Các công cụ khai phá dữ liệu chỉ ra (i) tương tác; hay (ii) mối quan hệ vốn có trong các biến. Người ta quan tâm đến (i) hiện nội dung thông tin; (ii) lấy đủ dữ liệu.

1. *Hiện nội dung thông tin.* Có nhiều cách biểu diễn nội dung thông tin. Hiện nội dung theo chiều thời gian. Các mối quan hệ trong dữ liệu được xem như *đặc trưng* của dữ liệu. Xu thế được xem là mối quan hệ được kiểm chứng, được gọi là *đặc trưng dự đoán*;
2. *Lấy đủ dữ liệu.* Lấy đủ dữ liệu không có nghĩa lấy nhiều. Một số chiều cần được lược bớt khi mô hình hóa. Dữ liệu nên lấy đủ theo mẫu yêu cầu, đủ để xác định mô hình;
3. *Các giá trị rỗng và giá trị thiếu.* Các giá trị rỗng không ứng với giá trị thế giới thực; giá trị thiếu do không thu thập được. Một vài công cụ khai phá yêu cầu thay giá trị này bằng giá trị mặc định, hay giá trị nội suy từ các giá trị lân cận;
4. *Hình dạng của bộ dữ liệu.* Không gian trạng thái là không gian gồm các điểm; điểm ứng với trạng thái bài toán. Thể hiện của dữ liệu là một trạng thái bài toán.

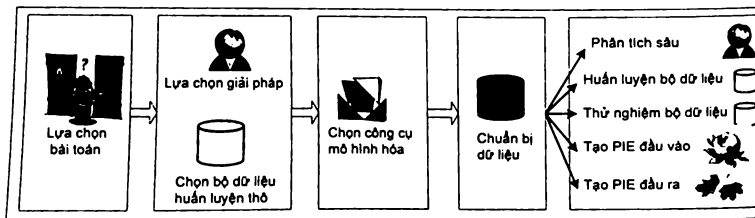


Hình 2.49. Điểm trong không gian pha 3D được thể hiện qua đa tạp

2.9. Quá trình chuẩn bị dữ liệu

Ở đây, việc chuẩn bị dữ liệu cho phân tích thống kê hay khai phá dữ liệu được xem như một quá trình nhiều bước. Ban đầu người ta cần xác định bài toán; và việc chuẩn bị dữ liệu được đặt trong ngữ cảnh của khám phá dữ liệu. Nếu không xác định được bài toán cần giải, sẽ khó xác định cách trích giá trị từ các hoạt động khai phá dữ liệu sau này.

Khai phá dữ liệu là công việc trên dữ liệu, ít nhiều phản ánh hoạt động của thế giới thực, các sự kiện, đối tượng. Việc chuẩn bị dữ liệu không phải quá trình tự động, sau này, khi trí tuệ nhân tạo phát triển, có thể công việc này sẽ tự động dần. Do các kỹ thuật chuẩn bị dữ liệu không thực hiện hoàn toàn tự động được, cần có sự can thiệp thủ công trong việc chuẩn bị dữ liệu.



Hình 2.50. Quá trình chuẩn bị dữ liệu

Người ta cũng đề cập nghệ thuật chuẩn bị dữ liệu, như thực tế thì nghệ thuật vẫn dựa vào các khía cạnh khách quan, cách thức các phần liên kết với nhau. Việc chuẩn bị dữ liệu có các khối (i) đầu vào; (ii) đầu ra; (iii) mô hình; và (iv) quyết định.

Như vậy cần giải thích về (i) đầu vào và đầu ra, và nhu cầu về chúng; (ii) mô hình tác động đến dữ liệu; (iii) các giai đoạn chuẩn bị dữ liệu và điều cần quyết định tại mỗi giai đoạn.

Mục đích cơ bản của chuẩn bị dữ liệu là xử lý và chuyển hóa dữ liệu thô, sao cho thể hiện được nội dung thông tin trong bộ dữ liệu, cho phép truy cập dễ dàng. Có hai quyết định chính (i) lời giải yêu cầu những gì; (ii) yêu cầu của công cụ khai phá.

Công việc bên trong của quá trình chuẩn bị được xem như hộp đen. Điều chú trọng ở đây là (i) vận hành của hộp đen; (ii) đầu ra từ hộp đen. Quá trình chuẩn bị là một trong ba bước chính của khai phá dữ liệu, tức (i) chuẩn bị dữ liệu; (ii) khảo sát dữ liệu; (iii) mô hình hóa dữ liệu. Mỗi bước có những đòi hỏi khác nhau trong quá trình chuẩn bị dữ liệu. Chúng tách biệt và kết thúc bước này mới sang bước sau.

2.9.1. Bước 1 về chuẩn bị dữ liệu

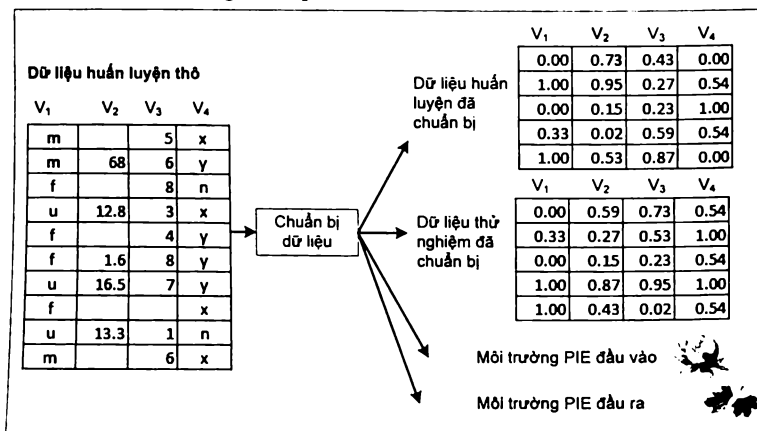
Lựa chọn bài toán là quá trình ra quyết định, tác động đến cả việc chọn lời giải, lẫn chọn dữ liệu. Lựa chọn công cụ mô hình hóa tiến hành theo bản chất của lời giải đã định và dữ liệu sẵn có.

Vài quyết định ban đầu liên quan đến cách chuẩn bị dữ liệu. Bản chất của bài toán xác định công cụ lựa chọn. Nếu có các quy tắc, người ta tuân theo. Người khám phá cần xác định cách phù hợp đối với việc chuẩn bị dữ liệu. Việc chuẩn bị có ít nhất bốn thành phần, như các đầu ra, (i) bộ dữ liệu huấn luyện; (ii) bộ dữ liệu thử nghiệm; (iii) khối đầu vào cho môi trường PIE; (iv) khối đầu ra cho PIE. Đó là các đầu ra cần thiết, với chức năng, mục đích và sử dụng nhất định. Chúng cần thiết, do bản chất của bộ dữ liệu được trích từ thế giới thực.

Bộ dữ liệu huấn luyện và bộ dữ liệu thử nghiệm cần cho mô hình hóa. Các khối môi trường PIE đầu vào cho phép mô hình ứng dụng cho bộ dữ liệu khác; khối môi trường PIE đầu ra dịch các câu trả lời của mô hình sang các giá trị đo được.

Công cụ khai phá dữ liệu có nhiệm vụ học về mối quan hệ vốn có giữa các biến trong bộ dữ liệu. Việc chuẩn bị bộ dữ liệu huấn luyện được thiết kế để công cụ mô hình hóa truy cập thông tin dễ dàng. Bộ dữ liệu không hoàn toàn phản ánh thế giới thực. Bản chất của quá trình đo là để nắm được độ không chắc chắn, vận xoắn, ồn. Mỗi công cụ mô hình hóa có cách khác nhau để diễn tả bản chất của các mối quan hệ giữa các biến. Tuy nhiên chúng khó xử lý các số đo, theo nghĩa tách được số đo thực và ồn. Đôi khi dữ liệu ồn cũng tạo nên mối quan hệ giữa các biến. Theo cách lý tưởng, công cụ mô hình hóa học cách đặc trưng mỗi quan hệ bên trong bộ dữ liệu, mà không học dữ liệu ồn.

Bộ dữ liệu thử nghiệm được dùng để kiểm việc tránh dữ liệu ồn. Cần phân biệt (i) bộ dữ liệu huấn luyện phát hiện mối quan hệ; (ii) bộ dữ liệu thử nghiệm phát hiện dữ liệu ồn.



Hình 2.51. Quá trình chuẩn bị dữ liệu chuyển dữ liệu thô thành bốn bộ dữ liệu

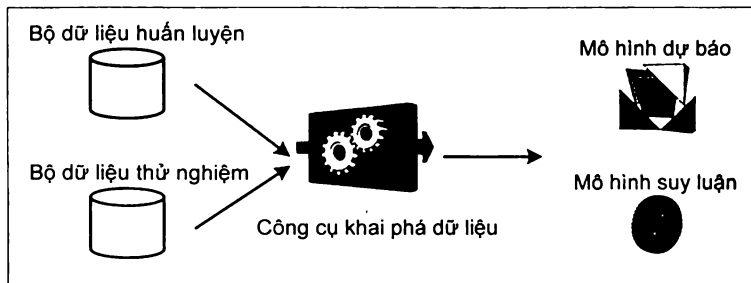
2.9.2. Bước 2 về khảo sát dữ liệu

Khai phá dữ liệu gồm cả khảo sát dữ liệu, tức dựng nên toàn cảnh mức cao để thấy được trong bộ dữ liệu có chứa nội dung gì. Vậy nên người khám phá nhìn sâu vào bản chất của dữ liệu.

Người ta khảo sát dữ liệu với nhiều kỹ thuật khác nhau.

2.9.3. Bước 3 về mô hình hóa dữ liệu

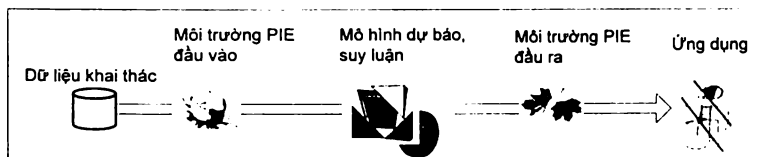
Người khai phá dữ liệu dùng công cụ mô hình hóa đã chọn để huấn luyện và thử nghiệm bộ dữ liệu, cho ra dự đoán mong muốn.



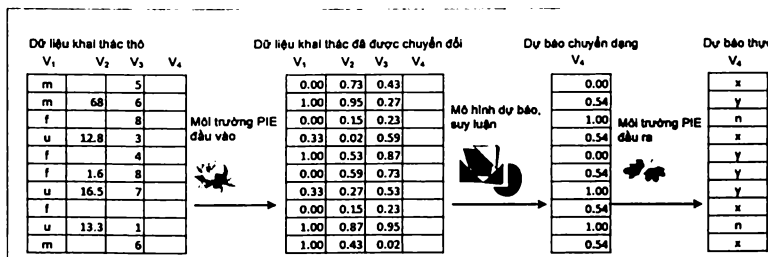
Hình 2.52. Khai phá mô hình dự đoán hay mô hình suy diễn

2.9.4. Sử dụng mô hình

Khi tạo được mô hình đáp ứng yêu cầu, để thực hành, người ta dùng dữ liệu thực, được gọi là *dữ liệu khai thác*.



Hình 2.53. Các bước sử dụng mô hình



Hình 2.54. Dự đoán và suy diễn thực sự với bộ dữ liệu thực

Trong hình trên, V_4 là biến phạm trù. Việc chuẩn bị dữ liệu chuyển tất cả các biến sang giá trị số vô hướng. Mô hình này dự

đoán kết quả ở dạng giá trị số. Tuy nhiên, dự đoán cần cho ở dạng phạm trù. Đây là mục đích của khối PIE đầu ra.

2.9.5. Công cụ mô hình hóa và chuẩn bị dữ liệu

Các công cụ khác nhau có ý nghĩa đối với công việc khác nhau. Khi xây dựng mô hình, trước tiên cần xác định (i) điều muốn phát hiện, và muốn hiểu biết; (ii) nguồn dữ liệu. Một số lớn các công cụ mô hình hóa hiện có đều có đặc trưng, mạnh, yếu khác nhau. Ngoài ra, những chương trình khai phá cần theo thuật toán, theo chủ đề khoa học, hay kinh doanh... Người khai phá dữ liệu cần quan tâm nhiều hơn, trước khi hiểu các yêu cầu của công cụ mô hình hóa, theo thuật ngữ của chuẩn bị dữ liệu, đặc biệt khi cơ sở dữ liệu đã tự động hóa, không có nhiều tương tác người dùng.

Tại pha thiết kế dự án, hay khi sử dụng trực tiếp công cụ mô hình hóa vạn năng, quan trọng là biết trước nhu cầu, sức mạnh, nhược điểm của mỗi công cụ.

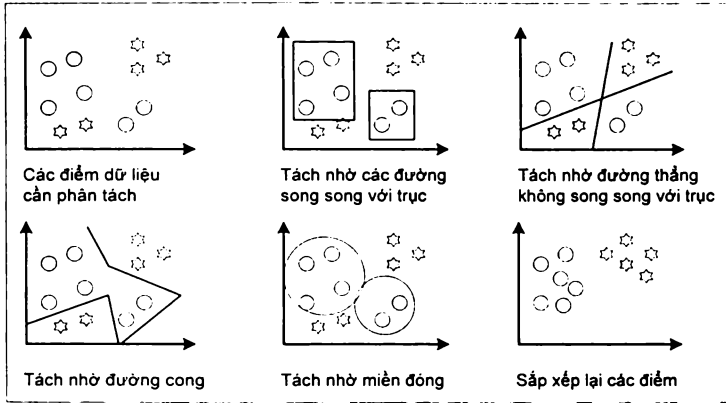
2.9.5.1. Công cụ mô hình hóa đối với chuẩn bị dữ liệu

Một nhân tố chính mà các công cụ khai phá tác động đến việc chuẩn bị dữ liệu là tính nhạy của công cụ đối với việc phân biệt dữ liệu số và dữ liệu phạm trù. Dữ liệu phạm trù nhạy với các giá trị thiếu, dù tính nhạy này bị hiểu sai nhiều.

Cách mà công cụ mô hình hóa lấy đặc trưng về mối quan hệ giữa các biến là phân đoạn dữ liệu, tức dữ liệu trong đoạn cụ thể liên quan đến đầu ra cụ thể. Các biến là (i) liên tục; (ii) rời rạc, nên một vài công cụ phân đoạn dữ liệu theo đoạn liên tục, hay rời rạc.

Hình sau cho thấy việc tách dữ liệu trong mô hình hóa bộ dữ liệu.

Các phương pháp cắt đa dạng cũng tương tự như cách mà các công cụ mô hình hóa chia dữ liệu. Hình minh họa cho thấy không gian có thể được cắt bằng các đường song song với trục tọa độ.



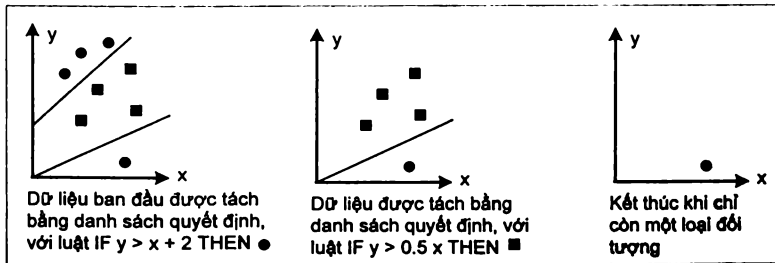
Hình 2.55. Mô hình hóa bộ dữ liệu

2.9.5.2. Cây quyết định

Cây quyết định sử dụng phương pháp kết hợp logic để xác định các miền của không gian trạng thái. Kết hợp logic này có thể được thể hiện ở dạng luật IF... THEN.

2.9.5.3. Danh sách quyết định

Danh sách quyết định cũng tạo nên luật IF... THEN, và về đồ họa, nó tương tự như cây quyết định. Tuy nhiên cây quyết định xét lực lượng con sang bên trái và bên phải nơi phân tách. Danh sách quyết định phát hiện luật cho phép đặc trưng cho một phần của lực lượng, rồi không xét ở những lần sau.



Hình 2.56. Danh sách quyết định suy luận các luật phù hợp của dữ liệu còn lại, cho đến khi xét đến tất cả các thể hiện

2.9.5.4. Mạng nơron

Mạng nơron cho phép không gian trạng thái được cắt thành các đoạn, với lát cắt không song song với trục tọa độ. Việc này được thực hiện khi mạng học với các trọng số tại mỗi nút mạng.

2.9.5.5. Chương trình tiến hóa

Chương trình tiến hóa là một dạng hồi quy, được biết như *hồi quy kí hiệu*. Nó phổ biến với quá trình phát hiện đẳng thức hồi quy, dùng trong phân tích thống kê, nhưng nó cũng cho phép khám phá mối quan hệ đặc biệt khó.

2.9.5.6. Mô hình hóa dữ liệu với công cụ

Thường người ta có công cụ thử nghiệm việc chuẩn bị dữ liệu, mô hình hóa, với tập dữ liệu thử nghiệm. Các công cụ này thể hiện các thí dụ minh họa, để người ta nắm được ý tưởng chính. Các công cụ ấy nói chung có ích đối với dữ liệu đã chuẩn bị theo các cách khác nhau.

2.9.5.7. Dự đoán và các luật

Lựa chọn công cụ sẽ tác động đến việc ứng dụng kĩ thuật nào lên dữ liệu chưa chuẩn bị. Tất cả các kĩ thuật mô tả ở đây đều sinh ra một hay hai dạng (i) dự đoán; (ii) luật. Các công cụ mô hình hóa kết thúc việc biểu diễn việc học, khi số dự đoán, phạm trù dự đoán, hay tập các luật có thể dùng được, để tách dữ liệu theo cách hữu hiệu.

Do việc dự đoán cần đến biến liên tục, nên công cụ phù hợp dùng cho mô hình là thứ có khả năng với dữ liệu liên tục. Công cụ ấy sinh ra loại đẳng thức, để biểu diễn mối quan hệ của các giá trị đầu vào với giá trị dự đoán.

Các phương pháp mô hình hóa phân đôi và liên tục có thể dùng cho nhiều mục đích, ngoài việc dự đoán. Liên quan đến dạng mô hình hóa suy luận, người ta xét tập các luật để khám phá lược đồ chung, thể hiện tình huống cụ thể.

2.9.5.8. Chọn kỹ thuật

Tác động của việc chọn công cụ mô hình hóa đến việc chuẩn bị dữ liệu là xác định bởi đặc trưng của công cụ. Cần xét công cụ đó làm việc với dữ liệu liên tục hay phạm trù. Do người ta thường làm việc với thuật toán điều chỉnh, nên mô hình thường giải quyết lớp bài toán cụ thể, chứ không giải vấn đề tổng quát.

2.9.5.9. Dữ liệu thiếu và công cụ mô hình

Dữ liệu thiếu có thể được cung cấp từ (i) giá trị mặc định; (ii) giá trị nội suy, để phù hợp với yêu cầu của thuật toán.

2.9.6. Các giai đoạn chuẩn bị dữ liệu

Chuẩn bị dữ liệu cần đến hai nhóm hoạt động (i) *thủ công*, tức không tự động, gồm hoạt động thủ tục ra quyết định về tiếp cận cần chọn; (ii) *tự động*, gồm các kỹ thuật, mã chương trình, quá trình và các điểm quyết định. Chuẩn bị dữ liệu được trình bày trong tám giai đoạn.

1. Truy cập dữ liệu;
2. Kiểm tra dữ liệu;
3. Tăng cường và làm giàu dữ liệu;
4. Tìm độ lệch của mẫu;
5. Xác định cấu trúc dữ liệu;
6. Xác định môi trường PIE;
7. Khảo sát dữ liệu;
8. Mô hình hóa dữ liệu.

2.9.6.1. Giai đoạn 1: Truy cập dữ liệu

Để truy cập dữ liệu, cần định vị chúng. Một số hệ thống giấu thuộc tính định danh, nên phải có công cụ để truy cập thuộc tính này. Về nguyên tắc, người dùng không được sử dụng ngôn ngữ

lập trình để trực tiếp truy cập tệp dữ liệu, mà phải qua ngôn ngữ hỏi dữ liệu.

Đối với kho dữ liệu, người ta không trực tiếp sử dụng dữ liệu tại đó. Thông thường đối với bài toán đã xác định, người ta sử dụng dữ liệu trong kho dữ liệu chuyên đề.

2.9.6.2. Giai đoạn 2: Kiểm tra dữ liệu

Việc kiểm tra dữ liệu¹ liên quan (i) nguồn cung cấp dữ liệu; (ii) chất lượng dữ liệu; (iii) số lượng dữ liệu. Để mô hình tốt, chuẩn bị dữ liệu cần đảm bảo dữ liệu đủ về số lượng, đảm bảo chất lượng phù hợp. Kiểm tra dữ liệu đảm bảo phương pháp luận để xác định trạng thái của bộ dữ liệu và ước lượng mức đủ để xây dựng mô hình.

Việc này cũng liên quan đến xem xét các mẫu nhỏ, đặt ra các câu hỏi để rà soát lại quá trình đã thực hiện.

2.9.6.3. Giai đoạn 3: Tăng cường và làm giàu dữ liệu

Dữ liệu không thừa; tuy nhiên người ta chỉ chuẩn bị vừa đủ. Khi bộ dữ liệu chưa đủ đáp ứng kì vọng, cần tăng cường dữ liệu. Thuật ngữ làm giàu dữ liệu mang ý tăng cường về số lượng và chất lượng dữ liệu.

Có nhiều cách tăng tính hữu dụng của dữ liệu. Chẳng hạn người ta sử dụng tỉ số giá và kết quả thu được, để xác định ngưỡng làm giàu dữ liệu. Tỉ số này (i) thể hiện cái nhìn sâu vào lĩnh vực; (ii) cho phép tiết kiệm thông tin tính trước về công cụ mô hình hóa.

2.9.6.4. Giai đoạn 4: Tìm độ lệch của mẫu

Độ lệch của mẫu thể hiện vấn đề cần giải quyết. Có vài phương pháp tự động phát hiện sự lệch của mẫu.

¹ Audit: kiểm tra.

Định nghĩa: Lấy mẫu là quá trình dùng phần nhỏ của bộ dữ liệu, để phần nhỏ này phản ánh chính xác mối quan hệ trong bộ dữ liệu.

Việc lấy mẫu được giới thiệu trong tài liệu về thống kê; tuy nhiên một số người dựng mô hình không tiếp cận tài liệu này, nên người ta chỉ dựa vào dữ liệu thu thập được để mô hình hóa.

2.9.6.5. Giai đoạn 5: Xác định cấu trúc dữ liệu

Cấu trúc thể hiện cách các biến trong bộ dữ liệu liên hệ với nhau. Qua cấu trúc dữ liệu, người ta có thể thấy (i) độ lệch; (ii) bản chất của cấu trúc; (iii) dữ liệu biến dạng...

Cấu trúc dữ liệu được nhìn nhận theo ba mức độ (i) *siêu cấu trúc*¹, phản ánh dữ liệu theo dạng toàn bộ dữ liệu; (ii) *cấu trúc lớn*, liên quan đến tạo dạng các biến; (iii) *cấu trúc nhỏ*, nói về cấu trúc tinh, mô tả cách các biến liên hệ với nhau.

2.9.6.6. Giai đoạn 6: Xác định môi trường PIE

Các giám sát chi tiết về dữ liệu nhằm (i) giúp xác định khả năng, tính cần thiết, điều chỉnh hay chuyển dạng dữ liệu; (ii) thiết lập kì vọng có lí của giải pháp thu được; (iii) xác định chất lượng chung, kiểm chứng về dữ liệu; (iv) bộc lộ sự thích hợp của dữ liệu đối với nhiệm vụ khám phá.

Nhiều hoạt động này đòi hỏi nhìn sâu vào ứng dụng, một cách thủ công, thay vì tự động qua công cụ công nghệ. Các quyết định được xác định qua một loạt các thao tác. Trong môi trường ra quyết định này, bộ dữ liệu có thể cho phép truy cập tự động từ máy tính.

- *Khía cạnh dữ liệu: mẫu biểu hiện.* Vấn đề ở đây là bao nhiêu dữ liệu cần cho mô hình hóa. Người ta thường theo nguyên lí *tất cả về dữ liệu, tất cả về thời gian*. Cần chỉ ra đặc trưng

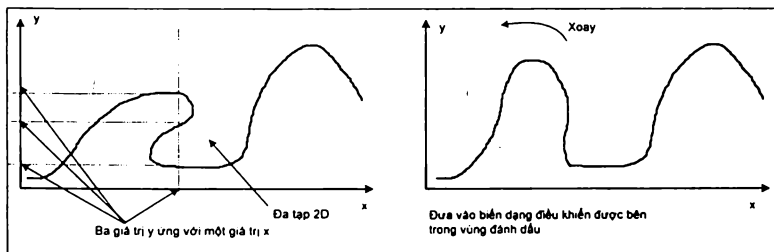
¹ Superstructure: siêu cấu trúc; macrostructure: cấu trúc lớn; Microstructure: cấu trúc nhỏ.

chính trong số các đặc trưng của dữ liệu. Việc xây dựng PIE là phức tạp, do nó thể hiện cấu trúc tinh của dữ liệu;

- *Khía cạnh dữ liệu: giá trị phạm trù.* Dữ liệu thuộc loại phạm trù được chuyển sang dạng số, để tiện cho quá trình làm việc tự động. Tuy nhiên, cuối cùng cần chuyển lại chúng về dạng ban đầu, tức phạm trù, để thích hợp với tự nhiên. Phương pháp chuyển sang dạng số tùy thuộc vào cấu trúc của bộ dữ liệu. Khi bộ dữ liệu gồm toàn dữ liệu phạm trù, người ta sẽ sử dụng tiếp cận xếp thứ tự các dữ liệu này. Điều quan trọng là không có phương pháp tự động cho công việc chuyển dữ liệu này, mà cần đến tri thức lĩnh vực của chuyên gia;
- *Khía cạnh dữ liệu: chuẩn hóa.* Chuẩn hóa liên quan đến (i) đưa dữ liệu về quy định chung; (ii) đưa các biến về cùng thang đo so sánh; (iii) tuân theo chuẩn về dữ liệu, chẳng hạn để giảm dư thừa dữ liệu. Bình thường bộ dữ liệu gồm nhiều bảng dữ liệu; nó phức tạp hơn một bảng, nên cần được chuẩn hóa;
- *Khía cạnh dữ liệu: giá trị thiếu và rỗng.* Không có kỹ thuật tự động để phân biệt (i) thiếu dữ liệu; (ii) dữ liệu rỗng. Tất cả những công cụ đều có phương tiện xử lý các giá trị thiếu, cho dù đã bỏ qua thể hiện dữ liệu chứa giá trị thiếu;
- *Khía cạnh dữ liệu: xâu chuyển dịch.* Xâu dữ liệu gắn với thời gian là dạng thường gặp trong thực tế. Điều này đòi hỏi người ta có các kỹ thuật xử lý riêng, ứng với thay đổi của thời gian;
- *Khía cạnh bộ dữ liệu: giảm độ rộng.* Dữ liệu được nhìn theo (i) số các thuộc tính, xem như *độ rộng* của bộ dữ liệu; (ii) số các dòng dữ liệu, tức các thể hiện dữ liệu, xem như *độ sâu* của bộ dữ liệu. Để tránh hiệu ứng bùng nổ dữ liệu, nhất thiết giảm chiều không gian trạng thái, hay ở đây là giảm chiều rộng của bộ dữ liệu. Vậy cần (i) chọn ra thuộc tính quan trọng; (ii) tổ hợp, tức gộp, các thuộc tính...;

- *Khía cạnh bộ dữ liệu: giảm độ sâu.* Giảm độ sâu của bộ dữ liệu là giảm số dòng dữ liệu, tức các thể hiện của đối tượng khác nhau. Như vậy người ta chỉ dùng tập con của bộ dữ liệu để mô hình hóa. Tập con của dữ liệu mô hình hóa phản ánh mối quan hệ vốn có trong bộ dữ liệu toàn thể;
- *Khía cạnh bộ dữ liệu và khảo sát dữ liệu: đa tạp tốt và chưa tốt.* Hình dạng của đa tạp dữ liệu cũng cần được quan tâm. Hình dạng đa tạp sẽ tiện cho xử lý và cải thiện dữ liệu. Việc khảo sát không liên quan đến việc xử lý dữ liệu, mà chỉ liên quan đến thông tin khám phá, trợ giúp quá trình mô hình hóa.

Hình vẽ cho thấy tìm kiếm đa tạp để quyết định xem có vấn đề hay không, để loại bỏ khi xử lý dữ liệu.



Hình 2.57. Biến dạng của đa tạp được điều khiển và được đưa vào dần dần, cho phép loại bỏ vấn đề

2.9.6.7. Giai đoạn 7: Khảo sát dữ liệu

Khảo sát dữ liệu xem xét và lập báo cáo về thuộc tính của đa tạp trong không gian trạng thái. Người khai phá dữ liệu quan tâm đến nhiều đặc trưng (i) mật độ tương đối của điểm trong không gian trạng thái; (ii) cụm xuất hiện tự nhiên; (iii) biến dạng không chính lí được; (iv) nơi xảy ra vấn đề.

2.9.6.8. Giai đoạn 8: Mô hình hóa dữ liệu

Mô hình hóa là công việc rộng, không đề cập chi tiết ở đây. Tuy nhiên người ta thấy cần thiết (i) cách xây dựng mô hình tương

tác với việc chuẩn bị dữ liệu; (ii) cách sử dụng dữ liệu đã chuẩn bị một cách hiệu quả.

2.10. Kết luận

Chương này đề cập dữ liệu sử dụng trong phân tích thống kê và khai phá dữ liệu. Đối với dữ liệu, cần (i) hiểu biết về dữ liệu; (ii) chuẩn bị dữ liệu cho quá trình phân tích thống kê và khai phá dữ liệu.

Dữ liệu được hiểu theo nhiều khía cạnh, theo dạng dữ liệu số, dữ liệu phạm trù. Người ta dùng các dạng thể hiện khác nhau của dữ liệu để làm phù hợp với quá trình phân tích thống kê hay khai phá dữ liệu. Việc hiển thị dữ liệu theo dạng thích hợp với người dùng là cần thiết trong tương tác người-máy của hệ thống thông tin. Độ đo về dữ liệu cho phép lượng hóa dữ liệu, so sánh các dữ liệu với nhau và chất lượng của dữ liệu.

Dữ liệu rời rạc hay dữ liệu liên tục đều có phân phối xác suất, và có đặc trưng phù hợp với bài toán cần giải. Chương này đã đề cập một số phân phối thông dụng. Các khái niệm liên quan đến phân bố thống kê là các kiến thức thuộc xác suất thống kê, nhưng cần thiết nhắc lại trong tài liệu, để tiện trình bày các giai đoạn chuẩn bị dữ liệu và xử lý dữ liệu về sau.

Việc chuẩn bị dữ liệu cho phân tích thống kê hay khai phá dữ liệu được xem như quá trình nhiều bước. Việc này dần dần được tự động hóa theo các bước của công nghệ dữ liệu.

Phương pháp, công cụ phân tích thống kê và khai phá dữ liệu

Trước khi tìm hiểu về thuật toán phù hợp trong các dự án khai phá dữ liệu, cần xét thuật toán mô hình hóa, như tập các công cụ khai phá dữ liệu. Thực hành khai phá dữ liệu sử dụng một số kĩ thuật, được phát triển và đặt trong bộ công cụ khai phá. Thời gian đầu, người ta sử dụng SQL và Perl và tùy biến. Nhiều chức năng như các đối tượng tách biệt trong bộ phần mềm khai phá. Hầu hết các dự án cần dùng vài công cụ không có trong bộ phần mềm khai phá dữ liệu. Các công cụ có ích thường là:

1. *Công cụ truy cập dữ liệu.* Công cụ này là ngôn ngữ SQL và các ngôn ngữ hỏi cơ sở dữ liệu khác;
2. *Công cụ tích hợp dữ liệu.* Công cụ trích rút, chuyển đổi và tải ETL¹ cho phép truy cập, thay đổi và tải dữ liệu từ các cấu trúc, định dạng khác nhau vào dạng đầu ra chung, tức các tệp phẳng, các cơ sở dữ liệu;
3. *Công cụ khám phá dữ liệu.* Đó là công cụ thống kê mô tả, đặc biệt các bảng tần suất, công cụ đào sâu, cắt lát...;
4. *Công cụ quản trị mô hình.* Các công cụ này gồm thư viện không gian khai phá dữ liệu, các mẫu, dự án;
5. *Công cụ phân tích mô hình hóa.* Có công cụ đánh giá mô hình, công cụ chọn đặc trưng;
6. *Công cụ linh tinh.* Đó là công cụ khác, gồm công cụ xử lý dữ liệu tại chỗ, công cụ triển khai nhanh, công cụ giám sát mô hình.

¹ ETL: Extract Transform Load: trích rút, chuyển đổi, tải.

Vấn đề là sử dụng công cụ một cách thích hợp sẽ giúp định ra các biến có ý nghĩa, tiện để ra quyết định nhanh.

3.1. Dữ liệu đặc trưng

3.1.1. Đặc trưng

Sau khi chuẩn bị bộ dữ liệu phân tích để mô hình hóa, cần lựa chọn với biến, hay các đặc trưng, dùng làm biến dự đoán. Quá trình chọn đặc trưng và chiến lược rất quan trọng, sau việc chuẩn bị dữ liệu để khai phá dữ liệu. Vấn đề chính trong khai phá dữ liệu trong bộ dữ liệu lớn, với nhiều biến dự đoán tiềm năng là *mối nguy về chiều*. Tài liệu của Richard Bellman, năm 1961, đã mô tả vấn đề về tăng các biến trong mô hình. Thêm một biến vào mô hình sẽ giúp (i) dự đoán tốt hơn trong mô hình hồi quy; (ii) phân biệt tốt hơn trong mô hình phân lớp. Vấn đề ở đây là khi thêm biến để phân tích, sự *hội tụ* trên các nghiệm trong hoặc quá trình tối thiểu hóa sai sót, hoặc quá trình học lặp, sẽ càng chậm. Lựa chọn đặc trưng nhằm giảm số biến trong mô hình, vậy sẽ giảm nguy cơ về chiều, nhờ việc loại bỏ các biến thừa hay không thích hợp, hay dữ liệu ổn. Điều này sẽ tác động tức khắc đối với việc (i) tăng tốc độ thực hiện thuật toán; (ii) nâng chất lượng dữ liệu; (iii) tăng năng lực dự đoán của thuật toán; (iv) làm kết quả dễ hiểu hơn.

Như vậy, đầu tiên là *danh sách ngắn* của các biến cần thiết cho khai phá dữ liệu. Danh sách tóm tắt này sẽ chỉ gồm các biến làm tăng ý nghĩa của năng lực dự đoán và tổng quát hóa khả năng của mô hình.

3.1.2. Biến đặc trưng

Định nghĩa: Đặc trưng là biến được chuyển hóa.

Các biến được xem như *thuộc tính*, *biến dự đoán*, hay *đặc trưng*¹. Nhưng trong vài lĩnh vực học máy, biến phân biệt với đặc trưng.

¹ Feature: đặc trưng.

Các máy học theo nhân, như máy vectơ tựa, chuyển các biến theo hàm toán học để liên hệ chúng với không gian lí thuyết bậc cao. Con người có thể hiểu không gian ba chiều, thậm chí bốn chiều. Còn toán học xác định không gian n chiều; mỗi chiều ứng với một chức năng toán học. Khi thực hiện chuyển hóa, các biến được chuyển hóa gọi là *đặc trưng*, không còn là biến nữa.

Định nghĩa: Ánh xạ¹ là quá trình tính toán để chuyển biến thành đặc trưng.

Mỗi biến được ánh xạ vào không gian chiều cao hơn, gọi là *siêu không gian*. Không gian này có thể xác định về toán học để có thể tách cụm đối với các điểm dữ liệu được ánh xạ, với mặt phẳng xác định theo các chiều. Vậy có khái niệm *siêu phẳng*, được cấu hình để tách nhiều cụm dữ liệu nhất trong bài toán phân lớp; như máy vectơ tựa đã làm. Cho dù các biến sẽ được ánh xạ vào các siêu phẳng nhờ thuật toán mô hình hóa, đôi khi người ta vẫn dùng lẫn thuật ngữ *biến* với *đặc trưng*.

3.1.3. Phương pháp xếp hạng đặc trưng

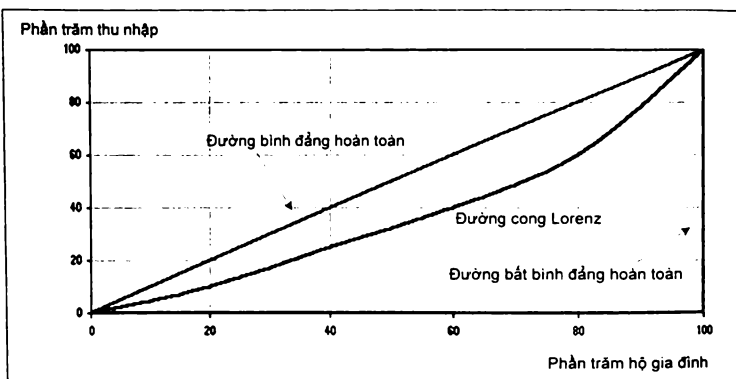
Có hai chiến lược chọn đặc trưng (i) phương pháp xếp hạng đặc trưng; (ii) chọn tập con tốt nhất.

Các phương pháp xếp hạng đặc trưng gồm (i) sử dụng các độ đo thống kê, như hệ số tương quan; (ii) sử dụng chỉ số Gini.

3.1.3.1. Chỉ số Gini

Chỉ số Gini dùng để lượng hóa sự không phẳng trong phân phối của biến, chẳng hạn như phân phối thu nhập theo các địa điểm. Lí thuyết cho việc này là sự sai khác giữa bình đẳng lí thuyết của lượng nào đó và giá trị thực trong khoảng của biến liên quan. Khái niệm này được Max O. Loen đưa ra năm 1905, để thể hiện phân phối bất bình đẳng giữa các địa điểm.

¹ Mapping: ánh xạ.



Hình 3.1. Đường cong Lorenz liên quan đến phân phối thu nhập của hộ gia đình

Phân phối lý thuyết của thu nhập theo hộ gia đình được ký hiệu bằng đường thẳng giữa hình. Tính bất bình đẳng thể hiện qua đường cong; đường này nằm sát đáy cho đến 80%, cho biết trong dân có ít người giàu và nhiều người nghèo.

Cerrado Gini tích hợp khái niệm của Lorenz vào năm 1912 để định lượng sự thay đổi về tần suất tương đối về giá trị thu nhập, theo khoảng dân số. Chẳng hạn chia trục hoành thành 10 đoạn, ứng với 10% số hộ, sẽ tính được số dân trong nhóm nhỏ đó, tức lượng hóa như tần suất tương đối. Tiếp cận chia ngăn này cho phép sử dụng tính toán theo tần suất, thay vì dùng phương pháp tích phân để tính diện tích dưới đường cong Lorenz tại mỗi điểm của trục hoành. Với tiếp cận này, người ta tính được *sai khác trung bình tương đối* RMD¹ của tất cả các giá trị xếp ngăn, tức tỉ số giữa tần suất của ngăn và tần suất trung bình của tất cả các ngăn, $RMD = MD / \text{trung bình số học}$; trong đó $MD = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|$, số dân lấy các giá trị y_i ; $i = 1..n$ là số dân.

Có thể dùng phương pháp này để hướng dẫn lựa chọn danh sách ngăn các biến, đưa vào thuật toán mô hình hóa. Chẳng hạn có

¹ RMD Relative mean difference: sai khác trung bình tương đối.

thể chọn các biến với điểm Gini lớn hơn 0.6. Nhược điểm của phương pháp này là nó kết hợp tác động của dữ liệu trong khoảng đã biết của một biến mà không phản ánh tác động kết hợp của tất cả các biến tương tác với biến này. Tuy nhiên đó cũng là vấn đề đối với hầu hết các phương pháp xếp hạng đặc trưng.

3.1.3.2. Phương pháp hai chiều

Phương pháp hai chiều khác, như thông tin tương hỗ, tính khoảng cách giữa phân phối nổi thực sự của đặc trưng X và Y , và tính phân phối nổi khi X độc lập với Y . *Phân phối nổi* là phân phối xác suất của các trường hợp mà sự kiện X và Y xảy ra đồng thời. Về hình thức, thông tin tương hỗ của hai biến ngẫu nhiên X và Y được xác định theo: $I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) * \log \frac{p(x, y)}{p_1(x) * p_2(y)}$;

trong đó $p(x, y)$ là phân phối xác suất nổi, và $p_1(x)$, $p_2(y)$ là hàm mật độ xác suất độc lập, hay xác suất lề, của X , Y . Đối với người làm thống kê, dễ dàng rút ra độ đo này. Ngược lại, người ta có thể chọn tiếp cận trực quan; có hai phương pháp (i) hồi quy từng bước; (ii) hồi quy bình phương tối thiểu từng phần.

3.1.3.3. Phương pháp nhiều chiều

Hồi quy tuyến tính từng bước: phương pháp khá lí tưởng là dùng hồi quy từng bước. Phương pháp thống kê cổ điển tính giá trị F đối với việc đưa vào thêm của mỗi biến trong hồi quy. Giá trị F tương đương với căn của giá trị t của Student, thể hiện hai mẫu khác nhau ra sao, trong đó một mẫu chứa biến và mẫu kia không chứa. Giá trị t được tính t như tỉ số của sai khác về trung bình mẫu và sai khác về độ lệch chuẩn; $F = \sqrt{t}$. Giá trị F nhạy với số biến tham gia vào từ số và mẫu số.

Hồi quy từng bước tính giá trị F đối với (i) trường hợp có biến tham gia; (ii) trường hợp không có biến tham gia, rồi so sánh nó với giá trị F *cực hạn*¹ để (i) lấy biến, tức lựa chọn từng bước tiến; (ii) loại

¹ Critical: cực hạn.

bỏ biến, tức lựa chọn từng bước lùi. Theo cách này thuật toán có thể chọn tập các biến thỏa mãn tiêu chí về giá trị F . Giả sử rằng số đến biến này là đủ thể hiện biến thiên tổng cộng trong biến đích, để dự đoán ở mức tin cậy đối với giá trị F , thường là 95%.

Nếu biến có dạng số, hay chuyển được sang dạng số, người ta có thể sử dụng hồi quy từng bước để chọn biến đối với thuật toán khai phá dữ liệu khác. Tuy nhiên cũng phiền phức. Hồi quy từng bước là thủ tục theo *tham số*, dựa trên cùng giả thiết đặc trưng trong các phương pháp thống kê cổ điển. Dù sao, hồi quy từng bước có thể dùng để cho thấy viễn cảnh về danh sách ngắn các biến. Có thể dùng phương pháp khác, rồi so sánh các danh sách. Về phương pháp này, đã giả sử về tính tuyến tính trong mối quan hệ giữa các biến với biến đích; nhưng thực tế sẽ không hoàn toàn như vậy.

Hồi quy bình phương tối thiểu từng phần: một biến thể phức tạp hơn của hồi quy từng bước cho phép giữ vết của tổng bộ phận cần bậc hai khi tính toán hồi quy. Các giá trị bộ phận này liên quan đến đóng góp của mỗi biến vào mô hình hồi quy. Bộ phần mềm thống kê *Statistica*¹ cho phép (i) phân tích dữ liệu; (ii) quản trị dữ liệu; (iii) thống kê; (iv) khai phá dữ liệu; (v) học máy; (vi) hiển thị dữ liệu. Phần mềm này cho báo cáo về hồi quy bình phương tối thiểu; từ đó cho phép lựa chọn đặc trưng.

Bảng 3.1. Đóng góp lẻ của sáu biến dự đoán đối với biến đích, tức lỗi tổng cộng

	Tăng, R^2 của Y
Biến 1	0.799
Biến 2	0.094
Biến 3	0.014
Biến 4	0
Biến 5	0
Biến 6	0

¹ Statistica do StatSoft phát triển, năm 2014.

Bảng sau cho biết thí dụ báo cáo phân tích về lỗi sản xuất. Rõ ràng là biến 1 và biến 2, và biến lẻ 3, đóng góp có ý nghĩa để dự đoán của mô hình, R^2 tổng là 0.9934. Trên cơ sở phân tích này, người ta có thể loại bỏ biến 4 đến biến 6 trong danh sách ngắn các biến.

Phân tích nhảy cảm: một vài thuật toán học máy, như mạng nơron, cho báo cáo đầu ra với tính toán trọng số gán cho mỗi biến, để tính đích *nhảy* ra sao khi thêm vào một biến. Các giá trị nhảy này tương tự như giá trị F, cho biết sự tham gia của mỗi biến trong hồi quy từng bước. Cả SPSS Clementine và Statistica đảm bảo báo cáo nhảy cảm cho mạng nơron tự động của chúng. Các giá trị nhảy này có thể dùng để phản ánh tập các biến tốt nhất tham gia vào mô hình. Một chiến lược là huấn luyện mạng nơron với các đặc tính mặc định và đưa tất cả các biến có mức nhảy cảm lớn hơn ngưỡng xác định, vào danh sách ngắn. Tiếp cận này kém chính xác hơn hồi quy từng bước, nhưng tập các biến của mạng nơron là tổng quát, theo ý khả năng của chúng để nắm được mối quan hệ phi tuyến tính.

3.1.3.4. Phương pháp phức tạp

Giới thiệu về phương pháp phức tạp: mạng tuyến tính từng mảnh sử dụng độ đo khoảng cách để gán trường hợp vào cụm thích hợp. Cụm này được xác định bằng phương pháp phân cụm. Hàm phân tách, gọi là *hàm cơ sở*, được xác định đối với mỗi cụm. Thuật toán xén cho phép loại bỏ các cụm ít quan trọng nhất, để mạng cô đọng hơn. Tiếp cận này là dạng phi tuyến của hồi quy tuyến tính từng bước.

Đường lười gà hồi quy thích nghi nhiều chiều: trong toán học, *đường lười gà*¹ là một hàm số được xác định theo từng phần bởi các hàm đa thức, và có độ mịn cao ở những nơi các phần đa thức kết

¹ Spline: lười gà.

nổi. Thuật toán lưới gà hồi quy thích nghi nhiều chiều MARS¹ do Friedman đề xuất năm 1991, cho phép giải bài toán phân lớp và hồi quy với nhiều đầu ra, tức nhiều biến đích. Tiếp cận này được xem như dạng hồi quy tuyến tính nhiều mảnh, theo lời giải cho miền cục bộ của lời giải tuyến tính tương tự. Mỗi miền cục bộ được diễn tả nhờ hàm cơ sở khác nhau.

Thuật toán MARS cũng có thể xem như dạng cây hồi quy, mà hàm phức tạp được tách thành các hàm cơ sở, mịn hơn ứng với các nhánh cây phân biệt. Statistica có thuật toán MARSplines, cho phép xén để lựa chọn đặc trưng. Thuật toán này chỉ lấy hàm cơ sở, và các biến dự đoán, để đóng góp có ý nghĩa cho dự đoán. Đầu ra của thuật toán này giữ lại chỉ các biến liên kết với hàm cơ sở trong lời giải cuối cùng, và xếp hạng chúng theo số lần chúng được sử dụng trong các phần khác nhau của mô hình.

3.1.4. Phương pháp lựa chọn tập con

Tiếp cận chọn tập con để chọn đặc trưng đánh giá tập con của đặc trưng có tác động ý nghĩa, như nhóm để dự đoán một biến đích. Các tiếp cận lựa chọn tập con thông dụng nhất dựa vào *chương trình bao*². Chương trình bao sử dụng thuật toán tìm kiếm để tìm trong không gian các đặc trưng có thể và đánh giá mỗi tập con, nhờ chạy mô hình trên tập con này. Vài chương trình bao thực hiện đánh giá với các tập con chọn ngẫu nhiên khác nhau, dùng phương pháp kiểm chứng chéo. Việc kiểm chứng chéo chia dữ liệu thành một số tập con đối với mỗi nhóm đặc trưng, và đánh giá mô hình được huấn luyện trên tất cả, trừ tập con. Trong bước lặp tiếp, tập con ngẫu nhiên khác được dùng để đánh giá.

Công cụ chọn đặc trưng của Statistica dễ sử dụng, tự động. Đặc biệt khối khai phá dữ liệu cho phép (i) tương tác, có chức năng chọn đặc trưng; (ii) không gian khai phá dữ liệu, có các điều

¹ MARS: Multivariate Adaptive Regression Splines: đường lưới gà hồi quy thích nghi nhiều chiều.

² Wrapper: vỏ, bao bọc; mục đích chính của nó là chương trình bao.

khuyến về chọn đặc trưng; (iii) phương thức khai phá dữ liệu, tự động việc chọn đặc trưng. Phần mềm Weka¹ cũng có khả năng của chương trình bao. Weka là một bộ phần mềm học máy được Đại học Waikato, New Zealand phát triển bằng Java, năm 2014. Nhưng cách khác về chọn đặc trưng theo chương trình bao là dùng RapidMiner. Nó là bộ khai phá dữ liệu mã mở GNU². RapidMiner, là một nền tảng phần mềm khoa học dữ liệu, được công ty RapidMiner phát triển năm 2006, cung cấp môi trường tích hợp cho việc học máy, học sâu, khai thác văn bản và phân tích tiên đoán. Nó có bốn phương pháp chọn đặc trưng (i) chọn lùi, dùng nhiều tập con; (ii) đánh trọng số đặc trưng, dùng láng giềng gần nhất; (iii) chọn đặc trưng theo chương trình bao; (iv) chọn đặc trưng tự động.

3.2. Phương pháp chọn mẫu

3.2.1. Phương pháp chọn mẫu trong nghiên cứu khoa học

Điều tra chọn mẫu có nghĩa là không tiến hành điều tra hết toàn bộ các đơn vị của tổng thể, mà chỉ điều tra trên một số đơn vị nhằm để tiết kiệm thời gian, công sức và chi phí. Từ những đặc điểm và tính chất của mẫu, có thể suy ra được đặc điểm và tính chất của cả tổng thể đó. Vấn đề quan trọng nhất là đảm bảo cho tổng thể mẫu phải có khả năng đại diện được cho tổng thể chung.

Từ tổng thể các dữ liệu, chọn ra n phần tử. Người ta gọi đó là *mẫu kích thước n* . Việc chọn mẫu theo cách (i) ngẫu nhiên; (ii) khách quan, mẫu được gọi là *mẫu ngẫu nhiên*. Hai cách lấy mẫu (i) *mẫu hoàn lại*³, tức phần tử vừa quan sát được trả về tổng thể, trước

¹ Weka: Waikato Environment for Knowledge Analysis: môi trường phân tích tri thức của Waikato.

² GNU: GNU được khởi đầu vào năm 1984, mục đích nhằm phát triển một hệ điều hành dạng Unix đầy đủ, đó là phần mềm tự do, hệ thống GNU.

³ With replacement: có thay thế, tức hoàn lại trước lần chọn mẫu sau; without replacement: không hoàn lại, tức lần chọn mẫu sau không có phần tử vừa được chọn.

khi lấy mẫu lần sau; (ii) *mẫu không hoàn lại*, tức phần tử vừa quan sát không được trả về tổng thể. Khi kích thước mẫu lớn, người ta không cần phân biệt việc hoàn lại hay không.

Mẫu được phân biệt (i) *mẫu định tính*, tức chỉ quan tâm đến phần tử có đặc trưng *A* nào đó; (ii) *mẫu định lượng*, tức chỉ quan tâm đến yếu tố định lượng của biến ngẫu nhiên.

3.2.1.1. Các bước chọn mẫu

Quá trình tổ chức điều tra chọn mẫu thường gồm sáu bước sau:

1. *Xác định tổng thể chung*. Người ta phải xác định rõ tổng thể chung, bởi cần chọn mẫu từ đó;
2. *Xác định khung chọn mẫu hay danh sách chọn mẫu*. Có thể chọn mẫu theo khuôn có sẵn, chẳng hạn các danh bạ điện thoại, danh sách liên lạc của nhóm người...;
3. *Lựa chọn phương pháp chọn mẫu*. Theo mục đích nghiên cứu, tầm quan trọng của công trình nghiên cứu, thời gian tiến hành nghiên cứu, kinh phí dành cho nghiên cứu, kĩ năng của nhóm nghiên cứu... mà người ta để quyết định (i) chọn phương pháp chọn mẫu xác suất; hay (ii) phi xác suất;
4. *Xác định quy mô mẫu*¹. Xác định quy mô mẫu thường dựa vào (i) yêu cầu về độ chính xác; (ii) đã có hay chưa có khung chọn mẫu; (iii) phương pháp thu thập dữ liệu; (iv) chi phí. Đối với mẫu xác suất, thường có công thức để tính cỡ mẫu; đối với mẫu phi xác suất: thường dựa vào kinh nghiệm và sự am hiểu về vấn đề nghiên cứu để chọn cỡ mẫu;
5. *Nhận diện được đơn vị mẫu trong thực tế*. Đối với mẫu xác suất, phải xác định rõ cách thức để chọn từng đơn vị trong tổng thể chung vào mẫu sao cho đảm bảo mọi đơn vị đều có khả năng được chọn như nhau.

¹ Sample size: kích thước mẫu.

6. *Kiểm tra quá trình chọn mẫu.* Các khía cạnh cần được kiểm tra (i) đơn vị trong mẫu có đúng đối tượng nghiên cứu; (ii) công tác của người trả lời; (iii) tỉ lệ mẫu theo đúng phương pháp lí thuyết.

3.2.1.2. Phương pháp chọn mẫu

Người ta có hai phương pháp chọn mẫu (i) chọn xác suất, ngẫu nhiên; (ii) chọn chủ quan, không ngẫu nhiên.

Phương pháp chọn mẫu ngẫu nhiên: Đây là phương pháp tốt nhất để có thể chọn ra một mẫu có khả năng đại biểu cho tổng thể. Vì có thể tính được sai số do chọn mẫu, nhờ đó có thể áp dụng được các phương pháp ước lượng thống kê, kiểm định giả thuyết thống kê trong xử lý dữ liệu để suy rộng kết quả trên mẫu cho tổng thể chung.

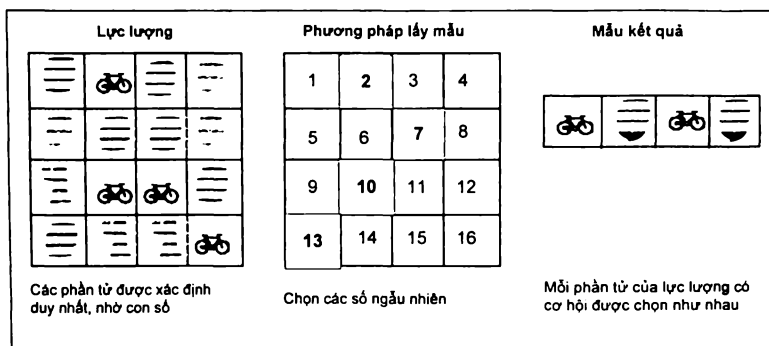
Định nghĩa: Chọn mẫu ngẫu nhiên¹, hay chọn mẫu xác suất, là phương pháp chọn mẫu mà khả năng được chọn vào tổng thể mẫu của tất cả các đơn vị của tổng thể đều như nhau.

Tuy nhiên ta khó áp dụng phương pháp này khi không xác định được danh sách cụ thể của tổng thể chung, chẳng hạn nghiên cứu trên tổng thể tiềm ẩn, tốn nhiều thời gian, chi phí, nhân lực cho việc thu thập dữ liệu khi đối tượng phân tán trên nhiều địa bàn cách xa nhau...

Tiếp theo là các phương pháp chọn mẫu ngẫu nhiên:

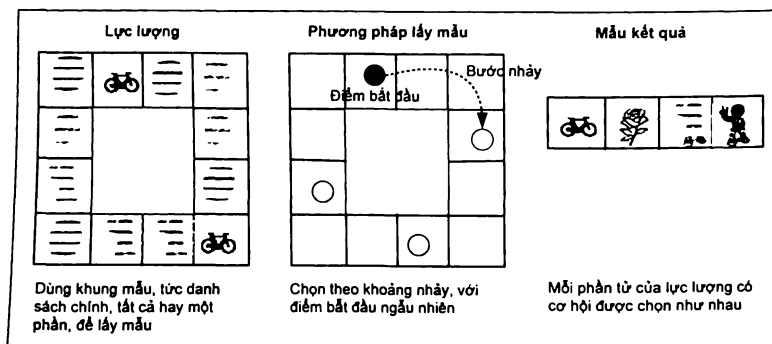
Chọn mẫu ngẫu nhiên đơn giản: trước tiên lập danh sách các đơn vị của tổng thể chung theo một trật tự nào đó: lập theo vần của tên, hoặc theo quy mô, hoặc theo địa chỉ..., sau đó đánh số thứ tự các đơn vị trong danh sách; rồi rút thăm, quay số, dùng bảng số ngẫu nhiên, hoặc dùng máy tính để chọn ra từng đơn vị trong tổng thể chung vào mẫu.

¹ Probability sampling methods: phương pháp chọn mẫu xác suất, ngẫu nhiên.



Hình 3.2. Chọn mẫu ngẫu nhiên đơn giản

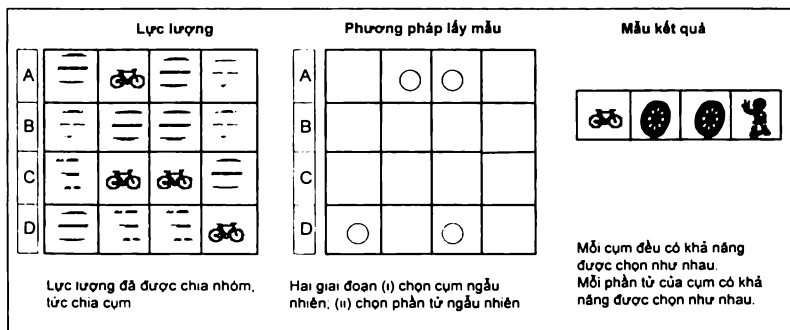
Người ta vận dụng phương pháp này khi các đơn vị của tổng thể chung không phân phối quá rộng về mặt địa lý, các đơn vị khá đồng đều nhau về đặc điểm đang nghiên cứu. Nó được áp dụng trong kiểm tra chất lượng sản phẩm trong các dây chuyền sản xuất hàng loạt.



Hình 3.3. Chọn mẫu ngẫu nhiên hệ thống

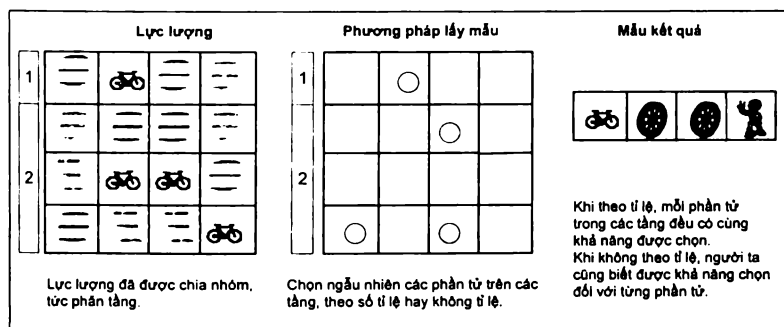
Chọn mẫu ngẫu nhiên hệ thống: Trước tiên, lập danh sách các đơn vị của tổng thể chung theo một trật tự quy ước nào đó, sau đó đánh số thứ tự các đơn vị trong danh sách. Đầu tiên chọn ngẫu nhiên một đơn vị trong danh sách; sau đó cứ cách đều k đơn vị lại chọn ra một đơn vị vào mẫu... cứ như thế cho đến khi chọn đủ số đơn vị của mẫu.

Chọn mẫu cả khối: Trước tiên, lập danh sách tổng thể chung theo từng khối; chẳng hạn làng, xã, phường, lượng sản phẩm sản xuất trong khoảng thời gian... Sau đó, chọn ngẫu nhiên một số khối và điều tra tất cả các đơn vị trong khối đã chọn.



Hình 3.4. Chọn mẫu cả khối

Người ta dùng phương pháp này khi không có sẵn danh sách đầy đủ của các đơn vị trong tổng thể cần nghiên cứu.



Hình 3.5. Chọn mẫu phân tầng

Chọn mẫu phân tầng: Trước tiên, phân chia tổng thể thành các nhóm theo tiêu chí hay nhiều tiêu chí liên quan đến mục đích nghiên cứu. Sau đó trong từng nhóm, người ta dùng cách chọn mẫu ngẫu nhiên đơn giản hay chọn mẫu hệ thống để chọn ra các đơn vị của mẫu.

Đối với chọn mẫu phân tầng, số đơn vị chọn ra ở mỗi nhóm có thể tuân theo tỉ lệ số đơn vị nhóm đó chiếm trong tổng thể, hoặc có thể không tuân theo tỉ lệ.

Chọn mẫu nhiều giai đoạn: Phương pháp này thường áp dụng đối với tổng thể chung có quy mô quá lớn và địa bàn nghiên cứu quá rộng. Việc chọn mẫu phải trải qua nhiều giai đoạn, cấp độ.

Trước tiên phân chia tổng thể chung thành các đơn vị cấp 1, rồi chọn các đơn vị mẫu cấp 1. Tiếp đến phân chia mỗi đơn vị mẫu cấp 1 thành các đơn vị cấp 2, rồi chọn các đơn vị mẫu cấp 2...

Trong mỗi cấp có thể áp dụng các cách chọn mẫu ngẫu nhiên đơn giản, chọn mẫu hệ thống, chọn mẫu phân tầng, chọn mẫu cả khối để chọn ra các đơn vị mẫu.

Phương pháp chọn mẫu chủ quan: Chẳng hạn tiến hành phỏng vấn các bà nội trợ tới mua hàng tại siêu thị tại một thời điểm nào đó; như vậy sẽ có rất nhiều bà nội trợ do không tới mua hàng tại thời điểm đó nên sẽ không có khả năng được chọn.

Định nghĩa: Chọn mẫu chủ quan, hay phi ngẫu nhiên, hay chọn mẫu phi xác suất, là phương pháp chọn mẫu mà các đơn vị trong tổng thể chung không có khả năng ngang nhau để được chọn vào mẫu nghiên cứu.

Việc chọn mẫu phi ngẫu nhiên hoàn toàn phụ thuộc vào kinh nghiệm và sự hiểu biết về tổng thể của người nghiên cứu nên kết quả điều tra thường mang tính chủ quan của người nghiên cứu. Mặt khác, ta không thể tính được sai số do chọn mẫu, do đó không thể áp dụng phương pháp ước lượng thống kê để suy rộng kết quả trên mẫu cho tổng thể chung

Sau đây là các phương pháp chọn mẫu chủ quan:

1. *Chọn mẫu thuận tiện:* Việc chọn mẫu này dựa trên sự thuận lợi hay dựa trên tính dễ tiếp cận của đối tượng, ở những nơi mà nhân viên điều tra có nhiều khả năng gặp được đối tượng. Chẳng hạn nhân viên điều tra có thể chặn bất cứ người nào mà họ gặp ở trung tâm thương mại, đường phố,

cửa hàng,... để xin thực hiện cuộc phỏng vấn. Lấy mẫu thuận tiện thường được dùng trong nghiên cứu khám phá, để xác định ý nghĩa thực tiễn của vấn đề nghiên cứu; hoặc để kiểm tra trước bảng câu hỏi nhằm hoàn chỉnh bảng; hoặc khi muốn ước lượng sơ bộ về vấn đề đang quan tâm mà không muốn mất nhiều thời gian và chi phí.

2. *Chọn mẫu phân đoán*: Đây là phương pháp mà phỏng vấn viên là người tự đưa ra phán đoán về đối tượng cần chọn vào mẫu. Như vậy tính đại diện của mẫu phụ thuộc nhiều vào kinh nghiệm và sự hiểu biết của người tổ chức việc điều tra và cả người đi thu thập dữ liệu. Chẳng hạn, nhân viên phỏng vấn được yêu cầu đến các trung tâm thương mại chọn các phụ nữ ưa nhìn để phỏng vấn. Điều này cũng đồng nghĩa với tính mờ của khái niệm *ưa nhìn*.
3. *Chọn mẫu định ngạch*: Đối với phương pháp chọn mẫu này, trước tiên ta tiến hành phân nhóm tổng thể theo một tiêu chí quan tâm, cũng giống như chọn mẫu ngẫu nhiên phân tầng, tuy nhiên sau đó ta lại dùng phương pháp chọn mẫu thuận tiện hay chọn mẫu phán đoán để chọn các đơn vị trong từng nhóm để tiến hành điều tra. Sự phân bố số đơn vị cần điều tra cho từng nhóm được chia hoàn toàn theo kinh nghiệm chủ quan của người nghiên cứu.

3.2.2. Giới thiệu phương pháp khởi động

Định nghĩa: Phương pháp khởi động, hay thúc đẩy khởi động¹ là tập hợp một số kĩ thuật phân tích dựa vào nguyên lí chọn mẫu có hoàn lại, để ước tính các thông số mà thống kê thông thường không giải được.

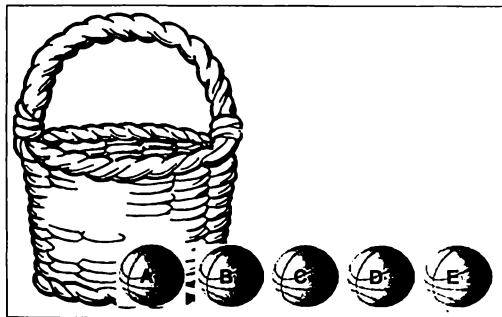
Phương pháp khởi động do nhà thống kê học Bradley Efron, đại học Stanford, Hoa Kỳ phát triển năm 1979, nhưng đến khi máy tính được sử dụng phổ biến thì phương pháp này mới trở thành phương pháp phổ biến trong phân tích thống kê và được ứng dụng rộng rãi trong rất nhiều lĩnh vực khoa học. Nó được xem là

¹ Bootstrap: khởi động, thúc đẩy khởi động.

phương pháp chuẩn trong phân tích thống kê và giải quyết được nhiều vấn đề mà trước đây tưởng như không giải được.

3.2.2.1. Tư tưởng chính của phương pháp khởi động

Phương pháp khởi động là phương pháp lấy mẫu có hoàn lại, tức một cá thể có thể xuất hiện nhiều lần trong một lần lấy mẫu. Giả sử ta có năm quan sát, với nhãn A, B, C, D và E trên năm quả bóng và bỏ tất cả chúng vào trong giỏ.



Hình 3.6. Thí dụ về các quan sát khác nhau

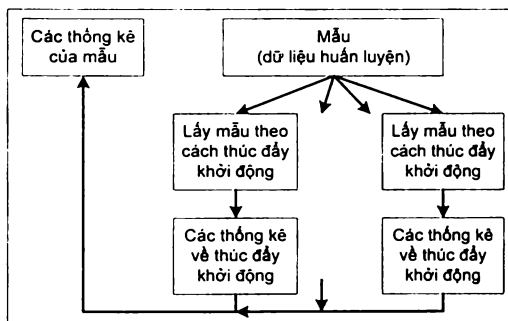
Từ năm quan sát này, ta lấy ra một quả bóng từ giỏ một cách ngẫu nhiên và ghi lại nhãn của chúng, sau đó bỏ lại quả bóng vừa bốc được vào giỏ và tiếp tục lấy ra một quả bóng một cách ngẫu nhiên, ghi lại nhãn của bóng và bỏ lại quả bóng vào trong giỏ và tiếp tục thực hiện việc lấy mẫu như vậy cho đến khi đủ số mẫu. Việc lấy mẫu này gọi là lấy mẫu có hoàn lại. Kết quả của việc lấy mẫu như trên có thể như sau, với kích thước mẫu là 10: $C, D, E, E, A, B, C, B, A, E$.

Lí do sử dụng phương pháp khởi động: Trong thực tế, từ một mẫu người ta chỉ có thể có được một số trung bình của mẫu, ta không biết được khoảng tin cậy cho số trung bình này hoặc không biết được phân phối của số trung bình ra sao. Thêm vào đó, thực tế không biết được hàm phân phối của tổng thể nên việc ước lượng các tham số đặc trưng thông kê rất khó khăn và thiếu chính xác. Phương pháp này có thể cung cấp nhiều thông tin chi tiết hơn

về phân phối của số trung bình, khoảng tin cậy cũng như xác suất của số trung bình dựa trên một mẫu duy nhất.

3.2.2.2. Các bước chính của phương pháp khởi động

1. Sinh ra các mẫu khởi động ngẫu nhiên có hoàn lại kích thước n từ tổng thể, từ mẫu ban đầu;
2. Tính các thông số thống kê đặc trưng cho của mẫu được sinh ra, tức trung bình, khoảng tin cậy, độ lệch chuẩn...;
3. Lặp lại bước 1 và bước 2 với số lần lớn, thường trên 1000;
4. Sử dụng các ước lượng thống kê của lấy mẫu khởi động đã tính ở bước 2 để đánh giá độ chính xác các ước lượng thống kê của mẫu ban đầu.



Hình 3.5. Các bước lấy mẫu khởi động

3.2.3. Giới thiệu các phương pháp nghiên cứu định tính

Phương pháp nghiên cứu định tính được sử dụng đầu tiên trong các nghiên cứu nhân chủng học.

Định nghĩa: Nghiên cứu khoa học là việc thu thập, phân tích và lí giải số liệu để giải quyết một vấn đề hay trả lời một câu hỏi.

Để có được những thông tin sâu, các nhà nhân chủng học đi đến sống ở các cộng đồng mà họ muốn nghiên cứu, họ thường sử dụng các kĩ thuật như phỏng vấn phi cấu trúc, thu thập lịch sử đời sống, thảo luận nhóm, nghiên cứu trường hợp để quan sát và tìm

hiểu những nguyên nhân chi phối hành vi ứng xử của người dân. Ngày nay, các kỹ thuật đó được sử dụng rộng rãi không chỉ trong phạm vi của nhân chủng học mà còn ở nhiều ngành, nhiều lĩnh vực khác nhau, đặc biệt là lĩnh vực y tế, xã hội học.

3.2.3.1. Phương pháp tiếp cận trong nghiên cứu định tính

Nghiên cứu định tính là một phương pháp tiếp cận nhằm tìm cách mô tả và phân tích đặc điểm văn hóa và hành vi của con người và của nhóm người từ quan điểm của nhà nghiên cứu.

Nghiên cứu định tính cung cấp thông tin toàn diện về các đặc điểm của môi trường xã hội nơi nghiên cứu được tiến hành. Đời sống xã hội được nhìn nhận như một chuỗi các sự kiện liên kết chặt chẽ với nhau mà cần được mô tả một cách đầy đủ để phản ánh được cuộc sống thực tế hàng ngày.

Nghiên cứu định tính dựa trên một chiến lược nghiên cứu linh hoạt và có tính biến chứng. Phương pháp này cho phép phát hiện những chủ đề quan trọng mà các nhà nghiên cứu có thể chưa bao quát được trước đó.

Trong nghiên cứu định tính, một số câu hỏi nghiên cứu và phương pháp thu thập thông tin được chuẩn bị trước, nhưng chúng có thể được điều chỉnh cho phù hợp khi những thông tin mới xuất hiện trong quá trình thu thập. Đó là một trong những khác biệt cơ bản giữa phương pháp định tính và phương pháp định lượng.

3.2.3.2. So sánh nghiên cứu định lượng và nghiên cứu định tính

Để thấy được vai trò của nghiên cứu định tính, trước hết xét vấn đề với nhược điểm của nghiên cứu định lượng. Liên quan đến các hạn chế đó, nghiên cứu định tính có thể giải quyết được.

Về nghiên cứu định lượng: Phương pháp nghiên cứu định lượng sử dụng một bảng hỏi đã chuẩn bị trước theo một cơ cấu nhất định cho mọi đối tượng nghiên cứu. Một ví dụ điển hình của

phương pháp này là điều tra KAP¹. Nghiên cứu KAP cho phép suy luận thống kê từ kết quả thu được ở các mẫu tương đối nhỏ ra quần thể lớn hơn; nó cũng cho phép đo lường và đánh giá mối liên quan giữa những biến số; tiến hành điều tra khá dễ và triển khai khá nhanh chóng và kết quả thu được từ các cuộc điều tra tốt có thể sử dụng để so sánh theo thời gian hoặc giữa các vùng.

Tuy nhiên KAP có một số nhược điểm và cần được sử dụng một cách thận trọng. Đáng lưu ý nhất là những sai số không do chọn mẫu, chẳng hạn người được hỏi trả lời không đúng các câu hỏi vì không nhớ hoặc do hiểu sai hoặc cố tình nói dối. Hai vấn đề nghiêm trọng nhất là:

1. *Thể hiện theo văn hóa vốn có.* Việc này xảy ra khi đối tượng phỏng vấn không hiểu câu hỏi đặt ra như ý định của nhà nghiên cứu mà lại hiểu khác đi và trả lời theo cách hiểu của họ;
2. *Sai số ngữ cảnh.* Những sai số ngữ cảnh là những yếu tố liên quan đến bản thân cuộc phỏng vấn. Phương pháp nghiên cứu định lượng giả định rằng hành vi và thái độ của con người không thay đổi theo ngữ cảnh. Tuy nhiên, câu trả lời của đối tượng có thể thay đổi phụ thuộc vào các ngữ cảnh khác nhau.

Nghiên cứu định tính: Nghiên cứu định tính cho phép các nghiên cứu viên hạn chế các sai số ngữ cảnh bằng cách sử dụng các kỹ thuật phỏng vấn và tạo ra một môi trường phỏng vấn mà trong đó đối tượng cảm thấy thoải mái nhất. Các phương pháp thu thập thông tin khác nhau đem lại thông tin khác nhau. Vì vậy trước khi quyết định sử dụng phương pháp nào cần phải xác định loại thông tin nào cần thiết nhất cho mục đích nghiên cứu. Các phương pháp nghiên cứu định tính và nghiên cứu định lượng có thể kết hợp để bổ sung lẫn cho nhau. Chẳng hạn:

¹ KAP: Knowledge – Attitude – Practice: Kiến thức - Thái độ - Thực hành.

- Nghiên cứu định tính có thể *trợ giúp* nghiên cứu định lượng bằng cách xác định các chủ đề phù hợp với phương pháp điều tra;
- Nghiên cứu định lượng có thể *trợ giúp* nghiên cứu định tính bằng cách khái quát hóa các phát hiện ra một mẫu lớn hơn hay nhận biết các nhóm cần nghiên cứu sâu;
- Nghiên cứu định tính có thể giúp *giải thích* các mối quan hệ giữa các biến số được phát hiện trong các nghiên cứu định lượng.

Bảng sau đây là lời khuyên để lựa chọn phương pháp nghiên cứu định lượng hay định tính.

Bảng 3.2. Nghiên cứu định lượng và nghiên cứu định tính

Sử dụng nghiên cứu định tính	Sử dụng nghiên cứu định lượng
Chủ đề nghiên cứu mới và chưa được xác định rõ.	Chủ đề nghiên cứu đã được xác định rõ và đã quen thuộc.
Nghiên cứu thăm dò, khi chưa nắm được những khái niệm và các biến số.	Khi những vấn đề cần đo lường khá nhỏ hay đã từng được giải quyết.
Khi cần thăm dò sâu, khi muốn tìm hiểu mối quan hệ giữa những khía cạnh đặc biệt của hành vi với ngữ cảnh rộng hơn.	Khi không cần thiết phải liên hệ những phát hiện với các bối cảnh xã hội hay văn hóa rộng hơn hay bối cảnh này đã được hiểu biết đầy đủ.
Khi cần tìm hiểu về ý nghĩa, nguyên nhân hơn là tần số.	Khi cần sự mô tả chi tiết bằng các con số cho một mẫu đại diện.
Khi cần có sự linh hoạt trong hướng nghiên cứu để phát hiện những vấn đề mới và khám phá sâu một chủ đề nào đó.	Khi khả năng tiến hành lại sự đo lường là quan trọng.
Nghiên cứu sâu và chi tiết những vấn đề được chọn lựa kĩ càng, những trường hợp hoặc các sự kiện.	Khi cần khái quát hóa và so sánh kết quả trong quần thể nghiên cứu.

Nghiên cứu định tính phù hợp với người khai phá dữ liệu với đặc điểm:

- Người nắm rõ thực địa;
- Người nắm rõ đối tượng nghiên cứu kể cả lịch sử, kinh tế, văn hóa, xã hội, nơi đối tượng sinh sống;
- Người được huấn luyện tốt, tức có kiến thức và kỹ năng tốt;
- Người có kinh nghiệm trong tiếp xúc với những người thuộc các thành phần xã hội khác nhau;
- Người kiên nhẫn và biết lắng nghe người khác.

3.2.3.3. Các phương pháp, kỹ thuật thu thập số liệu chủ yếu trong nghiên cứu định tính

Phòng vấn thường được dùng trong hai trường hợp (i) thể hiện chân dung nhân vật hoặc (ii) cung cấp thông tin về lĩnh vực mà người được phỏng vấn là chuyên gia hoặc có trách nhiệm trả lời.

Định nghĩa: Phòng vấn là một cuộc đối thoại có chủ đích, trong đó có người đặt câu hỏi và có người trả lời.

Phòng vấn báo chí có nhiều cách thực hiện (i) phỏng vấn trực tiếp, phóng viên và người được phỏng vấn gặp nhau mặt đối mặt; (ii) phỏng vấn bằng văn bản, tức người phỏng vấn chuẩn bị các câu hỏi bằng văn bản, người được phỏng vấn cũng trả lời bằng văn bản; (iii) phỏng vấn qua điện thoại hoặc các phương tiện liên lạc khác...

Phòng vấn không cấu trúc: Phương pháp phỏng vấn không cấu trúc, hay phi cấu trúc, là phương pháp được sử dụng rộng rãi nhất trong nghiên cứu xã hội học.

Định nghĩa: Phương pháp phỏng vấn không cấu trúc là phỏng vấn mà nghiên cứu viên phải nhớ một số chủ đề cần phỏng vấn và có thể sử dụng một danh mục chủ đề để khởi đầu trong khi phỏng vấn.

Nghiên cứu viên có thể chủ động thay đổi thứ tự của các chủ đề tùy theo hoàn cảnh phỏng vấn và câu trả lời của người được

phòng vấn. Phòng vấn không cấu trúc giống như nói chuyện, làm cho người được phỏng vấn cảm thấy thoải mái và cởi mở trả lời theo các chủ đề phỏng vấn. Điều cốt yếu quyết định sự thành bại của phỏng vấn không cấu trúc là khả năng đặt câu hỏi khơi gợi một cách có hiệu quả, tức là khả năng kích thích người trả lời cung cấp thêm thông tin.

Ưu điểm của phỏng vấn không cấu trúc là cho phép nghiên cứu viên linh hoạt thay đổi cấu trúc phỏng vấn tùy theo ngữ cảnh và đặc điểm của đối tượng. Nó đặc biệt có ích trong những trường hợp khi mà nghiên cứu viên cần phỏng vấn những người cung cấp thông tin nhiều lần, trong nhiều hoàn cảnh khác nhau. Phòng vấn không cấu trúc cũng hữu ích trong những trường hợp không thể sử dụng được phỏng vấn chính thức. Phòng vấn này đặc biệt hữu ích trong nghiên cứu các chủ đề nhạy cảm như tình dục, mãi dâm, ma túy hoặc HIV / AIDS ...

Nhược điểm của phỏng vấn không cấu trúc: Không có mẫu chuẩn bị sẵn nên mỗi cuộc phỏng vấn là một cuộc trò chuyện không lặp lại vì vậy rất khó hệ thống hoá các thông tin và phân tích số liệu.

Về phỏng vấn bán cấu trúc:

Định nghĩa: *Phòng vấn bán cấu trúc là phỏng vấn dựa theo danh mục các câu hỏi hoặc các chủ đề cần đề cập đến; thứ tự và cách đặt câu hỏi có thể tùy thuộc vào ngữ cảnh và đặc điểm của đối tượng phỏng vấn.*

Các loại phỏng vấn bán cấu trúc gồm:

Phòng vấn sâu. Được sử dụng để tìm hiểu thật sâu một chủ đề cụ thể, nhằm thu thập đến mức tối đa thông tin về chủ đề đang nghiên cứu. Phòng vấn sâu sử dụng bản hướng dẫn bán cấu trúc trên cơ sở những phỏng vấn thăm dò trước đó về chủ đề nghiên cứu để có thể biết được câu hỏi nào là phù hợp:

1. *Nghiên cứu trường hợp.* Nhằm thu thập thông tin toàn diện, có hệ thống và sâu về các trường hợp đang quan tâm.

Trường hợp ở đây có thể là một cá nhân, một sự kiện, một giai đoạn bệnh, một chương trình hay một cộng đồng. Nghiên cứu trường hợp đặc biệt cần thiết khi nghiên cứu viên cần có hiểu biết sâu về một số người, vấn đề và tình huống cụ thể, cũng như khi các trường hợp có nhiều thông tin hay mà có thể đem lại một cách nhìn sâu sắc về hiện tượng đang quan tâm;

2. *Lịch sử đời sống*. Thông tin về lịch sử đời sống của cá nhân thường được thu thập qua rất nhiều cuộc phỏng vấn kéo dài, thường là phỏng vấn bán cấu trúc và không cấu trúc.

Ưu điểm của phỏng vấn bán cấu trúc:

- Sử dụng bản hướng dẫn phỏng vấn sẽ tiết kiệm thời gian phỏng vấn;
- Danh mục các câu hỏi giúp xác định rõ những vấn đề cần thu thập thông tin nhưng vẫn cho phép độ linh hoạt cần thiết để thảo luận các vấn đề mới nảy sinh;
- Dễ dàng hệ thống hoá và phân tích các thông tin thu được.

Nhược điểm: Cần phải có thời gian để thăm dò trước chủ đề quan tâm để xác định chủ đề nghiên cứu và thiết kế câu hỏi phù hợp.

Phỏng vấn có cấu trúc: Thông tin thu được bằng phương pháp này có thể bao gồm cả các con số và các dữ liệu có thể đo đếm được. Các phương pháp này được coi là một bộ phận trong nghiên cứu định tính vì chúng giúp cho việc mô tả và phân tích các đặc điểm văn hóa và hành vi của đối tượng nghiên cứu.

Định nghĩa: *Phỏng vấn có cấu trúc, hay phỏng vấn hệ thống, là phương pháp phỏng vấn tất cả các đối tượng những câu hỏi như nhau.*

Các phương pháp này nhằm phát hiện và xác định rõ các phạm trù văn hóa thông qua sự tìm hiểu *những quy luật văn hóa* trong suy nghĩ của cá nhân, tìm hiểu xem họ nghĩ và biết gì về thế giới xung quanh họ và cách họ tổ chức các thông tin này.

- *Liệt kê tự do.* Tách biệt và xác định các phạm trù cụ thể. Nghiên cứu viên yêu cầu đối tượng liệt kê mọi thông tin mà họ có thể nghĩ tới trong một phạm trù cụ thể;
- *Phân lớp nhóm.* Phương pháp này tìm hiểu kiến thức của đối tượng về các phạm trù khác nhau và mối liên hệ giữa chúng;
- *Phân hạng sử dụng thang điểm.* Đây là phương pháp rất phổ biến trong khoa học xã hội. Các thang điểm thường được sử dụng để phân hạng các khoản mục trong một phạm trù nào đó. Thang điểm có thể là một dãy số, có thể là đồ thị.

Thảo luận nhóm: một điều cần lưu ý là đơn vị nghiên cứu và phân tích trong thảo luận nhóm sẽ là nhóm chứ không phải là cá nhân.

Thảo luận nhóm tập trung: Một nhóm tập trung thường bao gồm từ 6 đến 8 người có chung một số đặc điểm nhất định phù hợp với chủ đề cuộc thảo luận, chẳng hạn cùng một trình độ học vấn, cùng một độ tuổi, cùng một giới tính... Thảo luận nhóm tập trung thường được sử dụng để đánh giá các nhu cầu, các biện pháp can thiệp, thử nghiệm các ý tưởng hoặc chương trình mới, cải thiện chương trình hiện tại và thu thập các thông tin về một chủ đề nào đó phục vụ cho việc xây dựng bộ câu hỏi có cấu trúc...

Ưu điểm của phương pháp:

- Cung cấp một khối lượng thông tin đáng kể một cách nhanh chóng và rẻ hơn so với phỏng vấn cá nhân;
- Rất có giá trị trong việc tìm hiểu quan niệm, thái độ và hành vi của cộng đồng;
- Trợ giúp việc xác định những câu hỏi phù hợp cho phỏng vấn cá nhân.

Nhược điểm của phương pháp:

- Nghiên cứu viên khó kiểm soát động thái của quá trình thảo luận so với phỏng vấn cá nhân;

- Thảo luận nhóm tập trung không thể đưa ra tần suất phân phối của các quan niệm và hành vi trong cộng đồng;
- Kết quả thảo luận nhóm tập trung thường khó phân tích hơn so với phỏng vấn cá nhân;
- Số lượng vấn đề đặt ra trong thảo luận nhóm tập trung có thể ít hơn so với phỏng vấn cá nhân;
- Việc ghi chép lại thông tin và chi tiết của cuộc thảo luận nhóm tập trung rất khó, nhất là việc gỡ băng ghi âm.

Thảo luận nhóm không chính thức: Việc thảo luận nhóm không chính thức là thảo luận của các nhóm tự nhiên như nhóm thành viên gia đình, nhóm đàn ông uống trà trong quán, nhóm phụ nữ đi khám bệnh... Phương pháp này dùng kỹ thuật phỏng vấn bán cấu trúc hoặc phỏng vấn tự do. Phương pháp này dễ dàng thực hiện nhưng ít có tính hệ thống do đó khó sử dụng để so sánh giữa các nhóm. Phương pháp này có giá trị đối với các can thiệp đã được lập kế hoạch từ trước.

Phương pháp quan sát: Phương pháp phỏng vấn cung cấp các thông tin về quan niệm, thái độ, giá trị và hành vi tự thuật của đối tượng. Các phương pháp quan sát cung cấp thông tin về hành vi thực cho phép hiểu rõ hơn hành vi được nghiên cứu. Người ta có thể quan sát trực tiếp các hành vi thực tế hoặc có thể quan sát các dấu hiệu của hành vi. Đôi khi chỉ có thể quan sát gián tiếp dấu hiệu phản ánh hành vi.

Các hình thức quan sát:

- Quan sát tham gia/hoặc không tham gia;
- Quan sát công khai/hay bí mật;
- Giải thích rõ mục tiêu của quan sát/hoặc không nói rõ về mục đích thực của quan sát cho đối tượng bị quan sát biết;
- Quan sát một lần/quan sát lặp lại;
- Quan sát một hành vi/quan sát tổng thể;

- Quan sát thu thập số liệu định tính, mở và mô tả/quan sát thu thập số liệu định lượng dựa trên danh mục các điểm cần quan sát.

Phương pháp đi dạo là công cụ mô tả và chỉ ra vị trí và phân phối của các nguồn tài nguyên như đất, ruộng đồng, đồi núi, hồ, thềm thật vật, sông suối, nông trường, chuồng trại gia súc và các hình thái kinh tế, xã hội, văn hóa, tập quán cũng như các hoạt động thường nhật khác của cộng đồng... để xác định, giải thích có hay không sự ảnh hưởng, mối liên quan giữa các yếu tố trên với vấn đề sức khỏe hay những vấn đề khác mà cộng đồng đang đối mặt.

Định nghĩa: Phương pháp điều tra nhờ đi bộ là phương pháp điều tra, ghi chép... nhân lúc đi thực địa.

Trong phương pháp này, người nghiên cứu cùng với một vài thành viên của cộng đồng, thường là người am hiểu, sống lâu năm tại cộng đồng này, cùng đi bộ dạo quanh khu vực nghiên cứu để ghi chép, thảo luận, mô tả về sự phân phối các yếu tố địa lí, sinh cảnh, hiện tượng tự nhiên, xã hội... Những dữ liệu này rất quan trọng cùng với nguồn số liệu khác góp phần trả lời cho câu hỏi nghiên cứu.

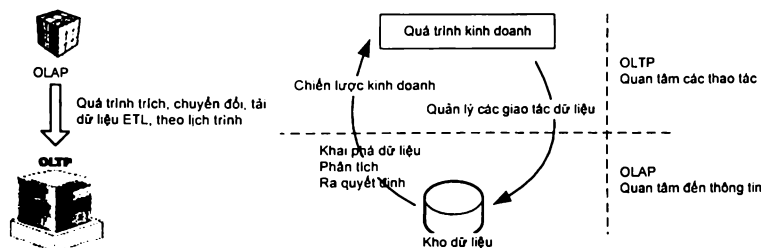
Nghiên cứu định tính là những nghiên cứu thu được các kết quả không sử dụng những công cụ đo lường, tính toán. Nói một cách cụ thể hơn, nghiên cứu định tính là những nghiên cứu tìm biết những đặc điểm, tính chất của đối tượng nghiên cứu cũng như những yếu tố ảnh hưởng đến suy nghĩ, hành vi của đối tượng nghiên cứu trong những hoàn cảnh cụ thể.

Trong khi nghiên cứu định lượng đi tìm trả lời cho câu hỏi bao nhiêu, mức nào, độ lớn của vấn đề, thì nghiên cứu định tính đi tìm trả lời cho câu hỏi nguyên nhân tại sao, ai thực hiện, cái gì xảy ra, như thế nào, ở đâu... giúp người ta đi sâu tìm hiểu một thực trạng hay một vấn đề nào đó.

3.3. Xử lý phân tích trực tuyến

3.3.1. Việc sử dụng OLAP

Công nghệ thông tin và truyền thông tập trung vào việc xây dựng hệ thống trợ giúp cho việc xử lý giao dịch của các tổ chức trong nhiều năm. Các hệ thống như vậy cần chịu được lỗi và đảm bảo đáp ứng kịp thời. Một giải pháp hiệu quả là xử lý phân tích trực tuyến OLAP¹, hướng vào môi trường cơ sở dữ liệu quan hệ, phân tán. Các phát triển mới nhất trong lĩnh vực này là sử dụng phần mềm ERP và SCM đối với các nhiệm vụ xử lý giao dịch, các ứng dụng CRM, và tích hợp với các công nghệ Web và intranet. Nhiều công cụ được tạo để phát triển các ứng dụng OLAP; máy chủ động trực tuyến Informix là thí dụ về công cụ hiệu quả.



Hình 3.8. OLTP và OLAP

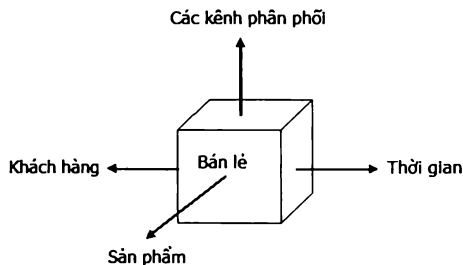
Truy cập đến dữ liệu thường cần thiết đối với các ứng dụng cả OLTP lẫn MSS. Tuy nhiên, cố gắng phục vụ cả hai loại yêu cầu này là khó. Do vậy, vài doanh nghiệp tạo nên cơ sở dữ liệu tách biệt, thành dạng OLAP và OLTP. Dạng OLAP dựa trên kho dữ liệu.

3.3.2. Vai trò của cơ sở dữ liệu đối với OLAP

Năm 2003, Golzales và Robinson cho thấy để làm việc phù hợp với OLAP, hệ quản trị cơ sở dữ liệu quan hệ cần được tối ưu,

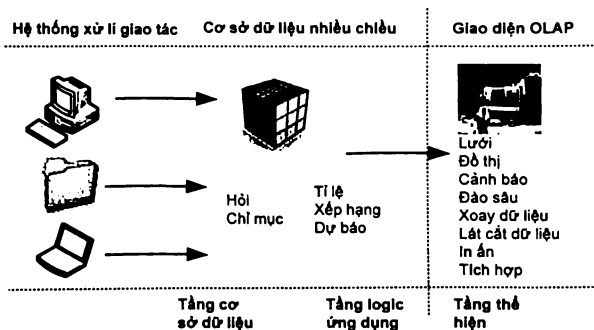
¹ OLAP: Online Analytical Processing; xử lý phân tích trực tuyến.

để trợ giúp OLAP, thay vì trực tiếp sử dụng đơn thuần các khối dữ liệu nhiều chiều.



Hình 3.9. Các chiều dữ liệu

Cơ sở dữ liệu cần được tích hợp với điều khiển tập trung, bền vững đối với dữ liệu nhiều chiều trong toàn xí nghiệp. Để cơ sở dữ liệu đoán trước các yêu cầu OLAP về dữ liệu mức cao trong tổ chức, các thư mục cơ sở dữ liệu cần có tập những đối tượng mức cao, có liên hệ trực tiếp với OLAP và các mô hình kinh doanh. Các đối tượng này ứng với các thực thể đang dùng, hợp chúng lại để tạo nên thực thể về chiều, như các thuộc tính, sự kiện, mối quan hệ, phân cấp và các chiều. Một khi xác định được các đối tượng mức cao, thông tin mới có thể được lưu trữ và quản lý như một phần của thư mục. Thực ra, việc quản lý siêu dữ liệu trở nên một phần của hệ quản trị cơ sở dữ liệu, cho phép tạo thành những đoán trước về OLAP.

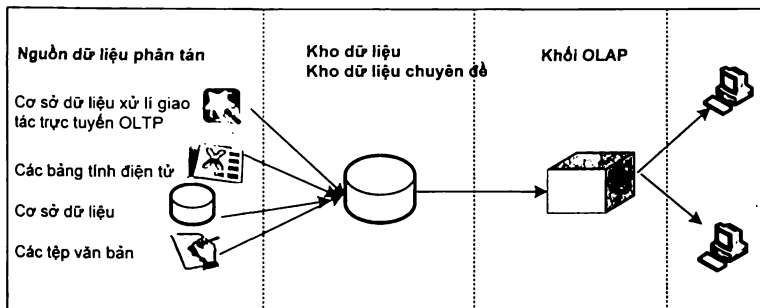


Hình 3.10. Phân tầng xử lý dữ liệu

3.2.3. Những hoạt động trực tuyến

Định nghĩa: Xử lý phân tích trực tuyến là các hoạt động người dùng trong hệ thống trực tuyến.

Thuật ngữ xử lý phân tích trực tuyến OLAP hướng đến nhiều hành động, thường do người dùng thực hiện trong các hệ thống trực tuyến. Người ta không xác định hoạt động nào là OLAP. Thông thường có thể kể ra các hoạt động, như (i) tạo sinh các câu hỏi, yêu cầu các báo cáo, đồ thị tùy biến; (ii) các phân tích thống kê; (iii) xây dựng các ứng dụng đa phương tiện và DSS. Ngoài ra còn có những hoạt động liên quan đến các hệ thống thông tin xí nghiệp hay điều hành và khai phá dữ liệu.



Hình 3.11. Hoạt động trực tuyến

Đặc biệt OLAP có khả năng mô hình hóa và hiển thị đối với bộ dữ liệu lớn. Bộ dữ liệu lớn có thể là cơ sở dữ liệu hay kho dữ liệu. OLAP khác với khai phá dữ liệu ở chỗ người dùng có thể hỏi các câu hỏi mở, chuyên biệt. Người ta, đặc biệt người phân tích, chạy các hệ thống OLAP. Họ vận hành OLAP, trong khi khai phá dữ liệu dùng để tìm kiếm các mối quan hệ, theo định hướng của người phân tích.

Nhìn chung, OLAP được trợ giúp từ phía kho dữ liệu, kho dữ liệu chuyên đề, kho dữ liệu nhiều chiều, với tập các công cụ OLAP. Các công cụ này có thể là (i) công cụ hỏi dữ liệu; (ii) bảng tính; (iii) công cụ khai phá dữ liệu; (iv) công cụ hiển thị dữ liệu.

3.3.4. Hỏi bằng ngôn ngữ SQL

Ngôn ngữ SQL¹ là ngôn ngữ chuẩn để truy cập, xử lý dữ liệu trong cơ sở dữ liệu quan hệ. SQL được dùng để truy cập trực tuyến đến cơ sở dữ liệu, các thao tác DBMS từ chương trình, và các chức năng quản trị cơ sở dữ liệu. Nó cũng được dùng để truy cập và xử lý dữ liệu của vài sản phẩm DBMS chính, như Oracle, DB2, Ingres.

Do SQL là ngôn ngữ phi thủ tục, thân thiện với người dùng, tại trạm đầu cuối người ta có thể hỏi dữ liệu. SQL có thể dùng trong ngôn ngữ lập trình chuẩn, nên nó thuận tiện đối với tích hợp phần mềm.

3.3.5. Các công cụ OLAP và tính chất

3.3.5.1. Các công cụ OLAP

Những công cụ phân tích và truy cập dữ liệu thường dùng là chưa đủ đối với OLAP. Trong OLAP, người ta dùng một lớp đặc biệt các công cụ, được xem như (i) *bên*² trợ giúp quyết định, thông minh doanh nghiệp, phân tích kinh doanh; (ii) bên truy cập dữ liệu; (iii) bên về cơ sở dữ liệu; và (iv) bên các hệ thống truy cập thông tin đồ họa. Các công cụ nhằm trợ giúp người dùng; bảng tính là một công cụ hiệu quả.

3.3.5.2. Tính chất của các công cụ OLAP

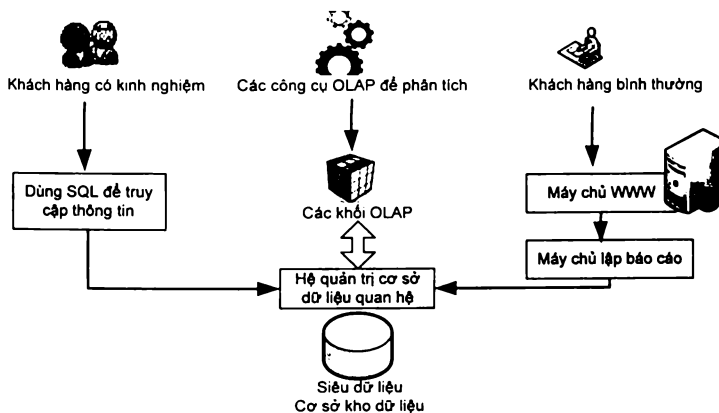
Các tính chất của các công cụ OLAP phân biệt chúng với các công cụ báo cáo được thiết kế cho các ứng dụng báo cáo OLTP truyền thống. Các đặc tính này do E. F. Codd đưa ra năm 1993. Đó là các luật đánh giá sản phẩm OLAP:

1. Khung nhìn khái niệm nhiều chiều;
2. Trong suốt;

¹ SQL: Structured Query Language: ngôn ngữ hỏi có cấu trúc.

² Front-end: máy phía trước.

3. Truy cập được;
4. Hiệu năng báo cáo bền vững;
5. Kiến trúc khách/chủ;
6. Chiều dữ liệu chung;
7. Quản lý ma trận thừa động;
8. Các thao tác với các chiều không hạn chế;
9. Xử lý dữ liệu trực quan;
10. Ra báo cáo mềm dẻo;
11. Không hạn chế các chiều và các mức gộp lớn.



Hình 3.12. Công cụ OLAP

3.3.6. Các loại xử lý của người phân tích

3.3.6.1. Các hoạt động phân tích

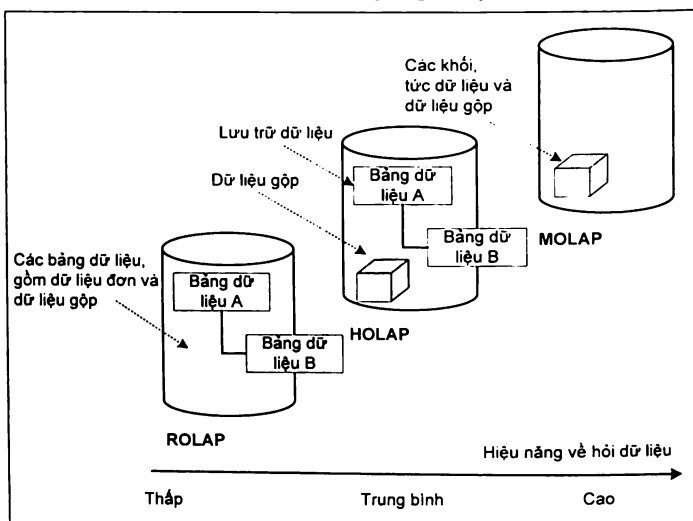
Người ta xác định bốn loại xử lý mà người phân tích sử dụng trong một tổ chức:

1. *Phân tích phạm trù*¹ là phân tích thống kê dựa trên dữ liệu lịch sử. Nó dựa vào giá trị rằng hiệu năng quá khứ là chỉ

¹ Categorical analysis: phân tích phạm trù.

số về tương lai. Đây là phân tích đầu tiên với trợ giúp của cơ sở dữ liệu theo giao dịch OLTP;

2. *Phân tích chú giải*¹ cũng dựa trên dữ liệu lịch sử, tạo điều kiện phân tích đào sâu vào dữ liệu. Phân tích đào sâu cho phép hỏi kĩ hơn về dữ liệu, để quyết định dữ liệu chi tiết có ảnh hưởng đến các giá trị suy diễn;
3. *Phân tích dự liệu*² cho phép người dùng thay đổi giá trị đơn, tăng tác động;
4. *Phân tích theo công thức* cho phép thay đổi nhiều biến.



Hình 3.13. MOLAP, ROLAP, HOLAP

OLAP được thực hiện với các kĩ thuật ROLAP, MOLAP và HOLAP³. (i) ROLAP truy cập dữ liệu trong kho dữ liệu theo mô hình quan hệ, để phân tích OLAP; (ii) MOLAP sử dụng các cơ sở dữ liệu nhiều chiều; (iii) HOLAP là kết hợp ROLAP và MOLAP,

¹ Exegetical analysis: phân tích chú giải.

² Contemplated analysis: phân tích dự liệu

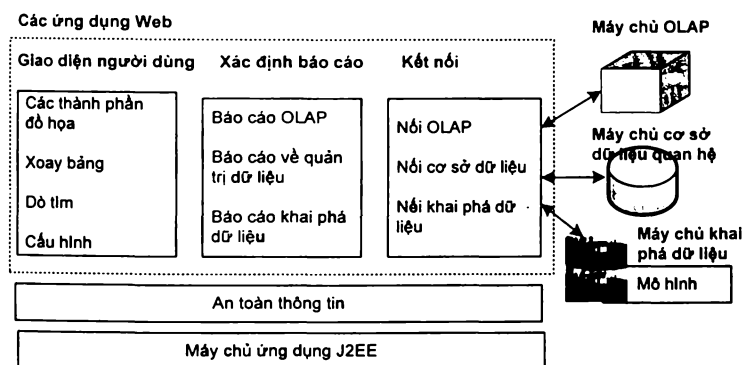
³ ROLAP: Relational OLAP, MOLAP Multidimensional OLAP, HOLAP: Hybrid OLAP: OLAP quan hệ, OLAP nhiều chiều, OLAP lai.

đặt một phần dữ liệu trong kho dữ liệu quan hệ, một phần trên cơ sở dữ liệu nhiều chiều, cân đối ưu điểm của cả hai kỹ thuật.

3.3.6.2. Nhà cung cấp OLAP

Các nhà cung cấp BI cố gắng để người dùng đầu cuối có khả năng cắt may các ứng dụng phân tích, để đáp ứng những nhu cầu tiến hóa về kinh doanh. Các nhà cung cấp gồm có Spotfire, Business Objects và Qliktech. Ngày nay có nhiều công cụ OLAP. Chúng có chung nhiều đặc trưng, nhưng cũng có những khác biệt lớn.

3.5.6.3. Các phép toán OLAP trong mô hình dữ liệu nhiều chiều



Hình 3.14. Ứng dụng OLAP

Trong mô hình nhiều chiều, dữ liệu được tổ chức thành các chiều, mỗi chiều chứa các mức trừu tượng do phân cấp khái niệm xác định. Tổ chức này đảm bảo mềm dẻo cho người dùng khi nhìn dữ liệu theo các hướng khác nhau. Một số phép toán về khối dữ liệu OLAP có khả năng hiển thị, có câu hỏi tương tác và phân tích dữ liệu. Tức là OLAP có môi trường người dùng để phân tích dữ liệu tương tác.

Các phép toán OLAP:

1. *Cuốn*: phép cuộn, hay phép *khoan*, thực hiện việc gộp trên dữ liệu khối, hoặc leo lên phân cấp khái niệm đối với một

chiều, hoặc giảm số chiều. Khi cuốn do giảm số chiều, một hay nhiều chiều bị loại khỏi khối dữ liệu. Cuốn có thể xóa chiều thời gian, cho kết quả là tổng sản phẩm theo địa điểm, chứ không còn theo địa điểm và thời gian nữa.

2. *Đào sâu*: phép đào sâu là một dạng cuốn. Nó quét từ dữ liệu ít chi tiết đến dữ liệu nhiều chi tiết. Đào sâu có thể thực hiện từng bước xuống phân cấp khái niệm đối với một chiều dữ liệu, hoặc đưa ra các chiều bổ sung. Chẳng hạn đi theo chiều thời gian để xét theo quý, rồi theo tháng... Do việc đào sâu thêm chi tiết đối với dữ liệu đã có, nó cho phép bổ sung chiều mới, chẳng hạn chiều "nhóm các khách hàng".
3. *Cắt lát và kẻ ô*: phép cắt lát chọn trên một chiều của khối dữ liệu, tạo nên một khối con. Chẳng hạn chọn thời gian = quý 1. Phép kẻ ô xác định khối con bằng cách chọn hai hay vài chiều. Chẳng hạn chọn (địa điểm = Hà Đông OR địa điểm = Hà Nội) AND thời gian = quý 1.
4. *Xoay*¹: phép xoay hiển thị khối con xoay quanh trục dữ liệu để có thể hiện khác của dữ liệu.
5. *Các phép OLAP khác*: vài hệ thống OLAP cung cấp các phép đào sâu, chẳng hạn đào chéo, đào qua² (i) đào chéo đặt câu hỏi trên nhiều bảng sự kiện; (ii) đào qua sử dụng tiện nghi ngôn ngữ SQL để đào qua mức thấp của khối dữ liệu đến bảng quan hệ đầu người dùng.

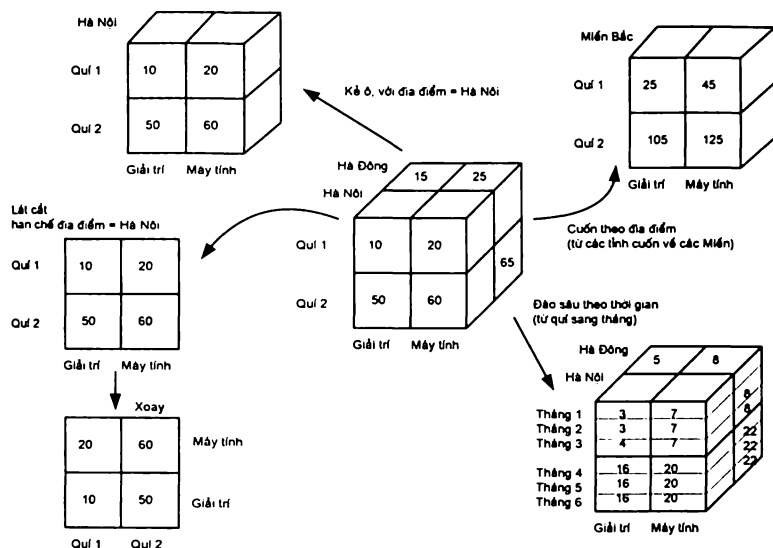
Các phép OLAP khác có thể sắp xếp hạng các mục theo danh sách, cũng như thực hiện tính toán phụ.

Định nghĩa: Cơ sở dữ liệu thống kê là hệ thống cơ sở dữ liệu được thiết kế để trợ giúp các ứng dụng thống kê.

¹ Pivot, Rotate: xoay.

² Drill across, Drill through: Đào qua.

³ Statistical database: cơ sở dữ liệu thống kê.



Hình 3.15. Quá trình thực hiện OLAP

So sánh các hệ thống OLAP với các cơ sở dữ liệu thống kê, người ta thấy: nhiều tính chất của hệ thống OLAP, như sử dụng mô hình dữ liệu nhiều chiều, kết hợp của độ đo với chiều, các phép toán cuốn, đào sâu cũng đã có trong cơ sở dữ liệu thống kê. Khác nhau giữa hai hệ thống nhiều khi là kĩ pháp, thuật ngữ.

OLAP và cơ sở dữ liệu thống kê cũng khác nhau. Cơ sở dữ liệu thống kê hướng ứng dụng kinh tế xã hội, OLAP có đích là ứng dụng nghiệp vụ. Khía cạnh riêng liên quan đến phân cấp khái niệm liên quan đến cơ sở dữ liệu thống kê. Không như cơ sở dữ liệu thống kê, OLAP được thiết kế để xử lí khối lượng lớn các dữ liệu một cách hiệu quả.

3.4. Một số lỗi khi khai phá dữ liệu

Người ta cho rằng công việc khai phá dữ liệu để trích xuất các mẫu hữu ích và lâu dài là một kĩ năng nghệ thuật cho là khoa học. Áp lực tăng lên về kết quả hấp dẫn của khai phá dữ liệu để đánh

lừa người ta. Vấn đề là làm thế nào để vẫn giữ được các nguyên tắc của phân tích thống kê, mà vẫn thu được kết quả vững bền.

Một số lỗi thông thường cần tránh được nêu dưới đây.

3.4.1. Giới thiệu về sai sót

Phương pháp nghiên cứu khoa học cho rằng phán xét tốt là do kinh nghiệm; tuy nhiên kinh nghiệm đôi khi thu được từ phán đoán xấu. Trong nhiều năm, từ năm 2000 đến nay, qua nhiều lần khai phá dữ liệu trong các lĩnh vực khác nhau, con người gặp không ít sai lầm. Việc hệ thống lại các lỗi và nguyên nhân gây nên lỗi là cần thiết.

3.4.2. Thiếu dữ liệu

Để thực sự nâng cao với phân tích thống kê, người ta cần có các trường hợp gán nhãn, tức có biến đầu ra. Chỉ với dữ liệu đầu vào, tất cả những thứ có thể thực hiện là (i) tìm kiếm tập con, thỏa mãn một số đặc trưng; (ii) phát hiện các chiều cho phép thể hiện biến thiên của dữ liệu, tức tìm thành phần chính.

Những kĩ thuật không giám sát ít nhiều hữu ích hơn mô hình dự đoán hoặc phân lớp tốt, có giám sát. Ngay cả với một biến đầu ra, các lớp và các kiểu dữ liệu thú vị nhất về quan sát thường là rất hiếm nếu theo độ đo thứ tự. Chẳng hạn, khoảng 1/10 người rủi ro về tín dụng trong vòng 2 năm, 1/100 người nhận quảng cáo sẽ mua hàng... Càng ít các sự kiện đáng quan tâm, càng cần nhiều dữ liệu quan sát, để đủ cho việc xây dựng mô hình đối với các trường hợp chưa rõ.

Chẳng hạn đối với dự án phát hiện gian lận trong hợp đồng chính phủ, các trường hợp gian lận thấy được là rất hiếm; mà nỗ lực vất vả ban đầu chỉ có thể làm giảm kích thước của tổng thể, trong khi sự kiện cần tìm lại quá nhỏ. Điều này cho thấy mô hình sử dụng để đảm bảo rằng các hợp đồng lớn hầu như chắc chắn không phải lừa đảo. Nhưng trong nhiều trường hợp gian lận đã biết, tốt đối với

người khai phá dữ liệu, nhưng tồi đối với người nộp thuế, cần có mô hình cho phép tự động đánh dấu các trường hợp nghi ngờ.

Không thể khai phá dữ liệu mà không có dữ liệu, nhưng không phải bất kì dữ liệu nào. Nhiều dự án khai phá dữ liệu thực hiện trên *dữ liệu tìm thấy*, không thu được kết quả thực nghiệm như thiết kế.

Một công ty tín dụng lí tưởng đã nhận ra điều này khi tìm kiếm ra quyết định về thị trường cho sản phẩm của mình trong lớp các ứng dụng mà trước đây thường bị xem là mạo hiểm. Có lẽ phát hành thẻ giới hạn thấp sẽ có lợi nhuận, và thậm chí giúp lôi kéo các ứng dụng. Nhưng công ty không có dữ liệu về các ứng dụng cho phép phân biệt ứng dụng nào cần thử. Họ dùng bộ lọc truyền thống, cho phép lọc trước những ứng dụng trong lần xem xét ban đầu. Vì vậy, công ty này đã cung cấp một lượng nhỏ tín dụng gần như ngẫu nhiên cho hàng ngàn ứng dụng nguy hiểm. Sau đó, họ xây dựng mô hình dự đoán lỗi, mà chỉ dựa vào thông tin ứng dụng ban đầu. Đầu tư lớn này trong việc *tạo ra* các dữ liệu phù hợp đã cho phép công ty mở rộng cơ sở ứng dụng của mình.

3.4.3. Chú trọng vào huấn luyện

Đối với kết quả ngoài mẫu, người ta thấy quan sát trực tiếp trên bảng dữ liệu là giải pháp tốt. Các nhà nghiên cứu tại trung tâm y tế MD Anderson ở Houston một thập kỷ trước đây sử dụng mạng nơron để phát hiện ung thư. Các kết quả ngoài mẫu cho kết quả khá tốt, mặc dù nếu phương pháp được huấn luyện vẫn tốt hơn. Họ cho rằng huấn luyện mạng sẽ cải thiện nó. Tuy nhiên nếu quá tập trung vào huấn luyện, họ mất hàng tuần, chứ không phải một ngày, mà kết quả huấn luyện chỉ khá lên không nhiều.

Đây là một trường hợp điển hình của việc *quá khít*¹, khi người ta luôn ám ảnh về việc đưa ra rất nhiều trường hợp huấn luyện, tập trung quá nhiều vào mô hình các trường hợp đặc thù, để

¹ Overfit: quá khít.

mong rằng sẽ phát hiện các trường hợp tương tự đối với dữ liệu chưa thấy. Việc học máy lúc đầu nghĩ học càng nhiều càng tốt, nhưng thực ra không thể đủ đáp ứng với tri thức chưa đầy đủ của các tình huống.

Cách quan trọng nhất để tránh quá khít là dự trữ dữ liệu. Nhưng do dữ liệu, đặc biệt các trường hợp quan tâm, cần sử dụng công cụ tái tạo mẫu, chẳng hạn như khởi động, xác nhận chéo... Thử nghiệm thống kê truyền thống có ý nghĩa là cách bảo vệ đơn giản, trong khi cấu trúc mô hình là một phần của quá trình tìm kiếm, mặc dù đó là mạnh nhất của độ đo *dựa vào phạt*¹, như *tiêu chuẩn thông tin Bayes* hay *độ dài mô tả tối thiểu*.

Liên quan đến tái tạo mẫu, nhiều thực nghiệm mô hình hóa được thực hiện, với mẫu dữ liệu khác nhau, để làm sáng tỏ sự phân phối của các kết quả. Nếu cần tách dữ liệu huấn luyện và dữ liệu đánh giá cùng lúc, kết quả chính xác đánh giá có thể chủ yếu do may mắn, tức hoặc tốt hoặc xấu. Khi tách, dùng mười cách khác nhau, huấn luyện với tập 90%, và đánh giá tập 10% ngoài mẫu, người ta thu được mười mức chính xác khác nhau. Giá trị trung bình của phân phối kết quả đánh giá có xu hướng chính xác hơn so với một thực nghiệm đơn, và nó cũng cho độ lệch chuẩn tin cậy.

Lưu ý rằng việc tái tạo mẫu đánh giá bất cứ điều gì được tổ chức như nhau, lặp lại suốt quá trình. Có thể thiết lập cấu trúc, các hạng thức của mô hình và tìm kiếm các giá trị tham số cho nó trên nhiều bộ dữ liệu, rồi các kết quả chính xác sẽ được áp dụng cho cấu trúc mô hình cố định đó. Hoặc có thể tự động hóa nhiều giai đoạn của quá trình, như phát hiện *lỗi*, lựa chọn đầu vào, phát hiện tương tác, và đưa toàn bộ quá trình vào trong vòng lặp tái tạo. Cuối cùng phân phối chính xác của quá trình đầy đủ sẽ lộ ra.

Người ta có nhiều mô hình chống chéo, cần chọn ra mô hình để ứng dụng. (i) Một tiếp cận là lựa chọn một mô hình đơn, có thể

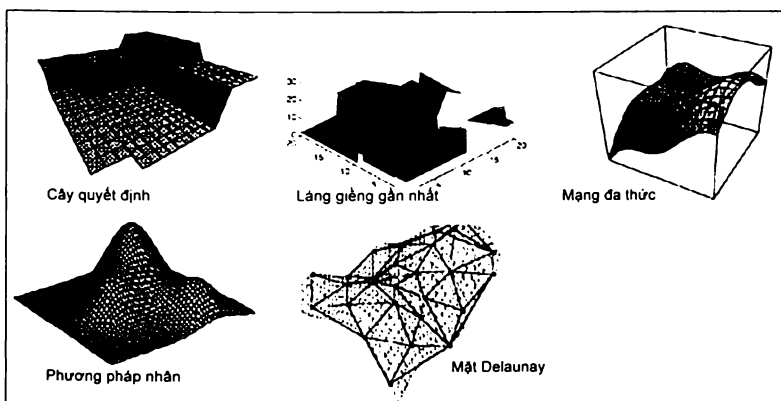
¹ Penalty - based: dựa vào phạt.

do cảm tính chứ chưa chính xác. (ii) Tiếp cận khác khai phá quá trình xây dựng mô hình, với tất cả dữ liệu và giả sử mô hình thu được sẽ kế thừa các thuộc tính chính xác nhờ các kiểm định chéo. (iii) Tiếp cận thứ ba dùng nhiều mô hình với nhau.

3.4.4. Chỉ tin cậy một kĩ thuật

Thường có nhiều giải pháp cho một vấn đề. Tuy nhiên có thể (i) sử dụng một giải pháp thích hợp; (ii) dùng công cụ nhiều giải pháp. Tiếp cận thứ hai chắc chắn cho kết quả tốt hơn.

Tuy nhiên, do nhiều lí do mà nhiều người khai phá dữ liệu chỉ chăm chú vào một kĩ thuật. Ít nhất, hãy xem xét một lượt, để ý đến phương pháp mới, đáp ứng yêu cầu. Chẳng hạn phương pháp hồi quy tuyến tính LR^1 hoặc phân tích khác biệt tuyến tính LDA^2 .



Hình 3.16. Ước lượng bề mặt của các thuật toán mô hình hóa

Chỉ sử dụng một phương pháp mô hình hóa đưa người ta đến tin tưởng, hay bài xích, nhằm thu được kết quả. Thông thường người ta đổ lỗi cho dữ liệu, nếu kết quả không tốt. Đối với kĩ thuật

¹ LR: Linear Regression: hồi quy tuyến tính.

² LDA: Linear Discriminant Analysis: phân tích khác biệt tuyến tính.

mô hình hóa cụ thể sẽ khác nhiều so với thực nghiệm của con người trên các dữ liệu khác nhau, khi mà khó đoán trước phương pháp này. Kinh nghiệm cho thấy với công cụ tốt, và dữ liệu tốt, người ta sẽ thu được kết quả; mà kiểm chứng với thuật toán quen thuộc sẽ thấy kết quả tin cậy.

Sự đa dạng thực sự của các thuật toán mô hình hóa là ít hơn nhiều hơn so với các dạng đã nêu ra. Nhưng có sự khác biệt thực sự trong cách xây dựng bề mặt nhờ *liên kết các điểm* của dữ liệu huấn luyện, như minh họa trong hình với năm phương pháp khác nhau (i) cây quyết định; (ii) mạng đa thức; (iii) tam giác Delaunay; (iv) nhân thích nghi; (v) láng giềng gần nhất. Cùng bài toán mà các mặt có đặc trưng khác nhau.

Người ta có thể đánh giá sai số tương đối giữa các phương pháp. Mỗi phương pháp đều phát huy điểm mạnh của nó đối với một lớp bài toán cụ thể. Theo các thực nghiệm, mạng nơron được đánh giá cao.

3.4.5. Đặt câu hỏi sai

Khi giải vấn đề, điều đầu tiên là xác định đích, để nhắm đến đúng mục tiêu. Điều này đã được minh chứng bằng một dự án tại Shannon Labs, của Daryl Pregibon, để phát hiện gian lận trong các cuộc gọi quốc tế. Thay vì sử dụng một phương pháp thông thường, cố gắng xây dựng một mô hình để phân biệt gian lận, ít dùng nhưng tốn kém, người ta đặt lại vấn đề. Từ các sự kiện bình thường, cho bất kỳ cuộc gọi nào đó, các nhà nghiên cứu đặc trưng mô hình gọi điện thoại bình thường đối với mỗi *tài khoản*, tức khách hàng, riêng biệt. Khi có cuộc gọi khởi hành từ những gì đã được mô hình bình thường cho tài khoản đó, cần bổ sung mức an ninh. Chẳng hạn một người thường được gọi từ một số ít quốc gia đặc biệt mỗi tuần, một thời gian ngắn, trong suốt các ngày trong tuần, cuộc gọi đến một khu vực khác nhau của thế giới vào cuối tuần sẽ chịu sự giám sát. Người ta tăng hiệu quả nhờ giảm thông

tin thanh toán lịch sử đối với khía cạnh chính, tạo ra một cơ chế cho các mức độ thích hợp của sự thích ứng theo thời gian, và thực hiện các mô hình trong thời gian thực đối với nguồn dữ liệu lớn. Tuy nhiên, chìa khóa để thành công vẫn là câu hỏi đúng về dữ liệu. Người ta đề xuất *chữ kí tài khoản*, đã đem lợi nhuận cho công ty. Đó là thí dụ ẩn tượng về khai phá dữ liệu lợi cho đầu tư ROI¹.

Ngay cả với mục tiêu dự án đúng, cần thiết có mô hình đích thích hợp. Máy tính cần cảm nhận được vấn đề như người dùng, để chia sẻ hàm đích đa nhân tố. Nhưng các nhà phân tích và bên cung cấp công cụ gần như luôn luôn sử dụng sai số bình phương như một tiêu chuẩn, chứ không phải là một thiết kế riêng cho bài toán này.

Mặc dù lỗi bình phương thường dùng có thể đặt một mô hình vào nơi hiệu quả, hàm được tối ưu có tác dụng triệt để về sự phù hợp của mô hình cuối cùng. Để xử lý số liệu tùy chỉnh tốt nhất, các nhà phân tích cần có thuật toán đa chiều mạnh mẽ, và đa hình thái. Ngay cả đối với việc tìm kiếm ngẫu nhiên, sử dụng hàm đích tùy chọn thường vẫn tốt hơn hàm không tùy biến.

3.4.6. *Chỉ tin vào dữ liệu*

Rút ra mô hình từ dữ liệu có ưu điểm là dựa vào dữ liệu hiện tại, không bị hạn chế bởi các giả thuyết cũ. Nhưng, để dữ liệu tự thể hiện, người ta không thu được nhiều tri thức. Kinh nghiệm đã cho các nhà phân tích thấy rằng những thứ quen thuộc với lĩnh vực thường quan trọng hơn so với các công nghệ, khi tìm giải pháp cho vấn đề.

Thông thường, không có gì *trong dữ liệu* là không có ý nghĩa; nhưng điều này không đúng. Bảng dữ liệu chứa hai biến về dữ liệu trường học, theo trung bình (i) chi phí và (ii) điểm SAT trung bình, từ năm 1994. Nhiệm vụ là mô hình hóa mối quan hệ của

¹ ROI: Return On Investment: lợi về đầu tư.

úng, để tư vấn giáo dục. Hình vẽ minh họa cách các mối quan hệ giữa hai biến là ý nghĩa. Xét các xếp hạng trên 4 người ta thấy mối quan hệ mạnh xảy ra một cách ngẫu nhiên chỉ 1/10,000 lần duy nhất, các dấu hiệu của mối quan hệ là ngược với những dự đoán. Đó là, để nâng cao vị thế, tức làm thấp xếp hạng SAT, đồ thị cho thấy cần giảm bớt học phí.

Bảng 3.3. Xếp hạng các địa phương theo đánh giá điểm SAT, của Hoa Kỳ

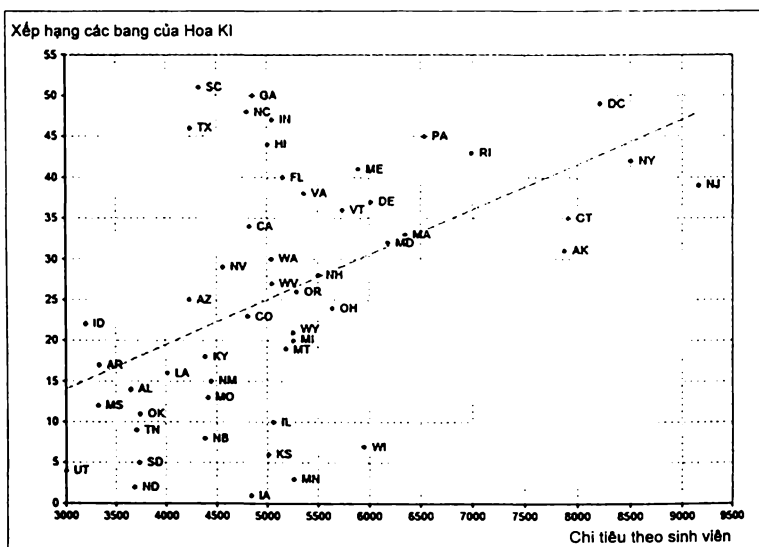
Tại Hoa Kỳ	Xếp hạng SAT	Chi phí (US\$)	Tại Hoa Kỳ	Xếp hạng SAT	Chi phí (US\$)
AK	31	7877	MO	13	4415
AL	14	3648	MS	12	3322
AR	17	3334	MT	19	5184
AZ	25	4231	NB	8	4381
CA	34	4826	NC	48	4802
CO	23	4809	ND	2	3685
CT	35	7914	NH	28	5504
DC	49	8210	NJ	39	9159
DE	37	6016	NM	15	4446
FL	40	5154	NV	29	4564
GA	50	4860	NY	42	8500
HI	44	5008	OH	24	5639
IA	1	4839	OK	11	3742
ID	22	3200	OR	26	5291
IL	10	5062	PA	45	6534

Tại Hoa Kỳ	Xếp hạng SAT	Chi phí (US\$)	Tại Hoa Kỳ	Xếp hạng SAT	Chi phí (US\$)
IN	47	5051	RI	43	6989
KS	6	5009	SC	51	4327
KY	18	4390	SD	5	3730
LA	16	4012	TN	9	3707
MA	33	6351	TX	46	4238
MD	32	6184	UT	4	2993
ME	41	5894	VA	38	5360
MI	20	5257	VT	36	5740
MN	3	5260	WA	30	5045

Người quan sát thí dụ này sẽ thường đề nghị bổ sung dữ liệu, chẳng hạn chi phí sinh hoạt của địa phương, hoặc phần trăm dân số ở thành thị hay nông thôn, để giúp giải thích những gì xảy ra. Nhưng vấn đề thực sự là một trong những tự lựa chọn. Các bang có SAT cao/chi phí thấp có cụm chính tại vùng Trung Tây, nơi mà thử nghiệm cần cho các trường đại học công không phải là SAT, mà là ACT. Chỉ những sinh viên tham vọng vào trường ngoài công lập sẽ gặp rắc rối vì theo thử nghiệm chuẩn hóa, và điểm trung bình kết quả của họ chắc chắn là cao hơn so với số đồng. Các biến khác trong cơ sở dữ liệu, trên thực tế sẽ làm cho các mô hình phức tạp hơn và có thể che khuất thực tế rằng các thông tin bên ngoài dữ liệu lại là quan trọng.

Thí dụ trên sử dụng điển hình *dữ liệu cơ hội*, hay tìm thấy. Nhưng ngay cả dữ liệu được tạo ra từ thực nghiệm được thiết kế cần thông tin ngoài. Một dự án của Bộ Quốc phòng Hoa Kỳ từ khi mới có mạng nơron đã cố gắng phân biệt hình ảnh trên không của khu rừng có và không có xe tăng. Người ta thu được hiệu suất

hoàn hảo sau khi huấn luyện mạng, và sau đó làm việc trên bộ điều kiện khác mẫu, thu thập đồng thời với dữ liệu mẫu. Hình ảnh mới được thu thập cho thấy mô hình hoạt động kém. Điều này khiến người ta tập trung nghiên cứu về đặc trưng điều khiển mô hình, và xác định được vai trò của điểm ảnh nền. Do người ta chụp ảnh xe tăng khi trời nắng; chụp ảnh không xe tăng lúc trời mây mù. Ngay cả việc tái tạo mẫu cũng không chống lại các lỗi này, như các lỗ hổng là vốn có trong các thực nghiệm tạo ra.



Hình 3.17. Xếp hạng địa phương theo trung bình của điểm SAT, so với mức chi tiêu của sinh viên, và ước lượng hồi quy bình phương tối thiểu của mối quan hệ này

Không một công nghệ mô hình hóa nào có thể tự chỉnh lý được các lỗi về dữ liệu, cần nghiên cứu cẩn thận về cách mô hình làm việc để phát hiện điểm yếu của nó.

3.4.7. Chấp nhận dò rỉ trong tương lai

Có người thường đánh giá các hệ thống đầu tư đầy hứa hẹn theo khả năng thực hiện. Có chuyên gia tư vấn, với nhiều tài liệu

hàn lâm, đã chuẩn bị mô hình mạng nơron một ngăn hàng, để dự đoán thay đổi lãi suất. Mô hình này chính xác đến 95%; đáng kinh ngạc. Hội đồng quản trị ngân hàng đã thận trọng tìm kiếm ý kiến thứ hai. Họ đã thấy rằng một phiên bản của các biến đầu ra đã vô tình được dùng như dữ liệu đầu vào.

Một mẹo nhỏ để đánh giá lỗ hổng trong hệ thống là nhìn kỹ vào biến đầu vào, khi nó hoạt động quá tốt. Chẳng hạn, trong dự án bảo hiểm, cố gắng xác định khách hàng của một câu lạc bộ ô tô mà là triển vọng tốt cho bảo hiểm, người ta đã thấy một đoạn mã đã có mặt khoảng 25% thời gian nhưng đã luôn gắn liền với bên mua bảo hiểm. Sau khi điều tra mở rộng, người ta thấy rằng mã là kiểu hủy bỏ bảo hiểm; đó là, nó thực sự đại diện cho thực tế rằng khoảng một phần tư số người mua hủy bỏ bảo hiểm của họ mỗi năm. Thường thì nơi có biến tốt nhất hóa ra lại là một nơi rò rỉ trong tương lai.

Nói chung, các kho dữ liệu được xây dựng để chứa các thông tin tốt nhất về khách hàng theo từng ngày. Như vậy, sẽ thuận tiện để rút ra thông tin trong quá trình học máy. Vì vậy, khi đang lưu trữ dữ liệu để khai phá trong tương lai, điều quan trọng là (i) ghi thời gian; (ii) lưu trữ định kỳ. Nếu không, các trạng thái thông tin thực tế được tạo lại sẽ trở nên khó và sẽ dẫn đến kết luận sai.

3.4.8. Giảm trường hợp phiến phức

Điểm ngoài lề và các điểm tựa có tác động, ảnh hưởng lớn đến kết quả tóm tắt và xu hướng đám mây hóa. Do vậy không nhất thiết loại bỏ các điểm này trong bộ dữ liệu. Qua việc phát hiện lỗ hổng tầng ozone tại Nam cực, người ta thấy dữ liệu ngoại lai cũng có giá trị trong quyết định.

Con người được xem là hoàn thiện, tuy nhiên con người không phải quá tốt, vậy nên cần chấp nhận các ngoại lệ, như đặc trưng vốn có.

Trong quá trình phân tích dữ liệu, người ta thường xử lý dữ liệu trong theo cách cho các biến hài hòa, theo một số biến chính. Phân tích tin cậy phụ thuộc nhiều vào chất lượng của dữ liệu. Những dữ liệu phù hợp nhau sẽ tiện thực hiện quá trình, nhưng không cho kết quả tốt; tuy nhiên nếu các dữ liệu có phần mâu thuẫn nhau sẽ gây khó khăn cho quá trình phân tích, nhưng đó mới đúng là thực tế.

3.4.9. Ngoại suy

Mô hình hóa đã nối các điểm, các trường hợp liên quan, để tạo nên các mối quan hệ, tạo thành không gian liên kết các đối tượng. Rõ ràng là các mô hình, đặc biệt mô hình phi tuyến là kém tin cậy bên ngoài các dữ liệu đã biết. Cũng vì vậy mà việc kiểm tra biên dữ liệu là nhiệm vụ cần thiết.

Nhưng có nhiều loại ngoại suy hay gây nguy hiểm, người ta có xu hướng học quá nhiều từ những kinh nghiệm đầu tiên, với một kĩ thuật hay vấn đề. Các giả thuyết được hình thành theo cách đơn giản hóa thế giới, thực ra không hợp lí đối với nhiều dữ liệu vốn đã mâu thuẫn. Như vậy việc tích lũy nhiều dữ liệu thực tế là quan trọng, hơn là sử dụng mô hình toán học để ngoại suy.

Việc ngoại suy với số chiều d nhỏ cũng gặp nguy hiểm như đối với d lớn. Về trực giác, kết quả ngoại suy với số chiều d nhỏ sẽ không có ích, nếu không có trường hợp phản tác dụng khi số chiều d lớn. Chẳng hạn trong thuật toán *láng giếng gần nhất*, giá trị đầu ra của các điểm gần nhất được xem như điểm mới. Khi số chiều d lớn, không có điểm điển hình thực sự gần nhau, tức có thể khoảng cách gần nhau, nhưng trên một chiều, có thể rất khác nhau.

Friedman, năm 1994, đã minh họa bốn thuộc tính của không gian với số chiều d lớn:

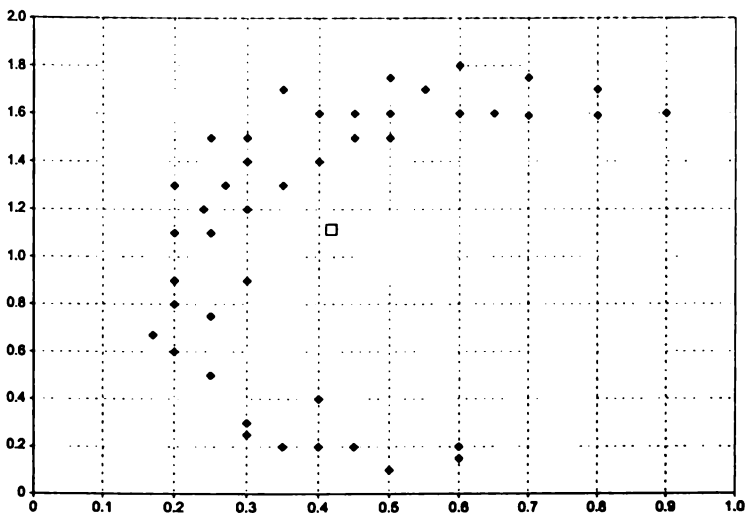
1. Kích thước của mẫu đạt được cùng mật độ tăng theo d , cấp mũ;

2. Bán kính bao phần dữ liệu đã cho tăng không tỉ lệ với d ;
3. Hầu hết các điểm là gần lề của không gian mẫu hơn là gần các điểm khác;
4. Hầu hết các điểm là ngoại lệ trong phép chiếu của nó.

Do người ta sử dụng kĩ thuật hiển thị, kĩ thuật mạnh và trực quan về quan hệ không gian, với d nhỏ, được điều chỉnh như với d lớn, người ta khuyên nên dùng các công cụ đơn giản nhất trong các bước đầu của bài toán. Việc này nhằm tách được các biến chính của bài toán và do vậy số chiều sẽ giảm.

3.4.10. Đáp ứng mọi yêu cầu

Không thể bắt hệ thống máy tính giải hết các bài toán con người đặt ra. Theo trí tuệ nhân tạo, hệ thống thông minh cũng phải có sai lầm và chỉ giải bài toán không quá khó. Điều này có nghĩa mô hình khai phá dữ liệu cần có một đầu ra ứng với trường hợp không giải được.



Hình 3.18. Bài toán hai chiều về lấy mẫu mà trung bình dữ liệu, ứng với hình vuông nhỏ trên hình, là bên ngoài biên của không gian được xác thực

Vấn đề là làm thế nào để biết mô hình có giá trị tại đâu, hay đã đủ dữ liệu cho vấn đề cần giải. (i) Tiếp cận thứ nhất, đơn giản, là xét điểm mới có nằm bên ngoài biên không, hay ngoài chiều, của bài toán được huấn luyện. Đặc biệt trong không gian có số chiều d lớn, khối lượng điểm trong không gian được lấp đầy chỉ là phần nhỏ so với khối lượng trong hình chữ nhật do biến một chiều tạo nên. Trong bài toán thực tế, các đầu vào là rất xa do độc lập lẫn nhau, vậy nên phần không gian bị chiếm là rất nhỏ, cho dù d nhỏ. (ii) Tiếp cận thứ hai, ít dùng và phức tạp, là tính bao lỗi của mẫu.

Hình ở đây mô tả bài toán hai chiều trong ứng dụng hàng không. Ràng buộc căn bản trên các giá trị nổi của các biến vật lí, như độ cao, tốc độ... khiến các điểm xa nhau, do *độc lập và phân phối đồng nhất*¹.

3.4.11. Mẫu ngẫu nhiên

Các trường hợp thuận lợi đối với vấn đề khai phá dữ liệu là rất hiếm, và thách thức về phân tích tựa như tìm kim. Bài toán sẽ dễ giải nếu tỉ lệ giữa đối tượng cần tìm và đối tượng cần tìm không quá lớn. Để bài toán có mức cân đối đủ để giải, người ta cần tránh hai thái cực (i) *dưới mẫu*², tức bỏ đi các trường hợp chung nhất; (ii) *trên mẫu*³, tức đã lặp lại nhiều trường hợp đặc thù, hiếm thấy.

Một công ty tiếp thị trực tiếp tại Maryland, Hoa Kỳ đã có quá nhiều trường hợp không đáp ứng bài toán ủng hộ từ thiện; tức có khoảng 99% cho mô hình cây quyết định. Để cân bằng dữ liệu, công ty giữ lại các trường hợp đáp ứng, và lấy một phần mười của các trường hợp không đáp ứng trong khoảng một triệu trường hợp, cho đến khi có 10,000 trường hợp. Khi khai phá dữ liệu,

¹ Iid: independent and identically distributed: độc lập và phân phối đồng nhất.

² Undersample: dưới mẫu.

³ Oversample: trên mẫu.

người ta thấy tất cả đều làm từ thiện; đó là kết quả sai. Người ta thấy chiến thuật tráo trộn dữ liệu trước khi lấy mẫu có tác dụng đối với bài toán này.

Lưu ý rằng, có thể cần phải xác định đơn vị lấy mẫu với mức chi tiết, độ mịn cao hơn trường hợp. Có nghĩa vài trường hợp có thể cần đi kèm với nhau và không bao giờ được tách ra trên các tập con dữ liệu. Chẳng hạn với các dữ liệu y tế, mỗi ca bệnh thường đại diện cho một lần thăm khám của bác sĩ. Trong kết quả dự đoán, gần như chắc chắn phải giữ tất cả các bản ghi từ ứng với một bệnh nhân với nhau và không áp dụng việc huấn luyện để đánh giá.

Cuối cùng, mẫu phân tầng hầu như luôn giúp người khai phá dữ liệu tránh rắc rối. Nên cẩn thận xem xét các biến cần thể hiện trong mỗi tập con dữ liệu và lấy mẫu một cách riêng biệt. Chẳng hạn lấy 10% các trường hợp màu đỏ và 10% màu xanh cho mỗi mẫu, thay vì lấy đơn giản 10% tổng thể. Việc phân tầng tốt là cách khá chính xác dự đoán chính trị; chẳng hạn người ta thực hiện với số ít phòng vận, khoảng 1,000.

3.4.12. Tin mô hình tốt nhất

Người ta có phát biểu rằng *tất cả các mô hình đều sai, ngoài một số là hữu ích*. Khi có nhiều mô hình, người ta có nhiều lựa chọn, nhưng cũng tốn kém, mà chưa chắc cho kết quả tốt hơn hẳn. Việc thay đổi mô hình cũng đồng nghĩa với việc đảm bảo dữ liệu đầu vào khác đi. Do vậy người ta luôn có xu hướng lựa chọn một mô hình tốt nhất để sử dụng.

Tuy nhiên, mỗi mô hình chỉ đáp ứng một số bài toán, khái niệm mô hình tốt nhất không bền. Như vậy, cũng cần thử nghiệm với vài mô hình. Mô hình được lựa chọn đã được kiểm nghiệm, so sánh kết quả với các thuật toán đã quen thuộc.

3.4.13. Định hướng

Thành công được cải thiện nhờ học hỏi, mà tốt nhất là do các thực nghiệm, kinh nghiệm, nhiều khi là được từ những sai lầm. Đương nhiên dù phải tập thực hành, vấp ngã, nhưng không ai muốn mắc nhiều sai lầm cả.

Sau đây là một số *con đường* dẫn đến thành công:

1. *Kiên trì.* Tấn công nhiều lần bài toán khai phá dữ liệu, từ các góc độ khác nhau. Tự động hóa các bước cần thiết, đặc biệt để có thể thực hiện các thử nghiệm tái tạo mẫu. Kiểm tra công việc từ phía ngoài, để khách quan. Một ý tưởng trong việc này là thuê người phá vỡ mô hình của mình, để tiện phát hiện lỗi;
2. *Thái độ.* Cần lạc quan, luôn sẵn sàng thực hiện việc có thể, nhất là khi làm việc theo nhóm;
3. *Làm việc theo nhóm.* Doanh nghiệp và các chuyên gia thống kê phải hợp tác chặt chẽ để tiến độ tốt nhất. Ý mong muốn thành công của dự án theo nhóm là không giống nhau. Có những cộng tác thụ động tích cực. Mỗi bên cung cấp dữ liệu hay thông tin nhạy cảm đều có mục đích riêng. Các mong muốn riêng cần được tạo thành ý muốn chung, để tiến cùng một hướng;
4. *Khiêm tốn.* Học hỏi từ những người khác đòi hỏi phải có đức tính khiêm tốn. Khi người khai phá dữ liệu thăm khách hàng, ít nhất họ thấy được chủ đề bài toán. Tuy người ta biết nhiều về phân tích thống kê và tránh được sai lầm, nhưng cần khiêm tốn về khả năng công nghệ. Trong công tác quản trị, nhà quản lí tuy không thạo công nghệ, nhưng là người ra quyết định và kí séc chuyển khoản. Luôn yêu cầu chuyên gia thực hiện từng bước, nhìn nhận tổng thể và đặt các câu hỏi đúng.

3.5. Kết luận

Việc chọn dữ liệu đặc trưng và chọn mẫu dữ liệu cũng liên quan đến việc hiểu và chuẩn bị dữ liệu, đã đề cập trong chương trước. Việc xử lý dữ liệu trực tuyến là biện pháp thông dụng trong khám phá tri thức, gắn với cơ sở dữ liệu. Các sai sót trong quá trình khai phá dữ liệu là các lỗi thường gặp, mà người khai phá dữ liệu cần tránh.

Chương này giới thiệu một số công cụ liên quan đến khai phá dữ liệu. Đó là (i) chọn các đặc trưng của dữ liệu, đề cập phương pháp xếp hạng đặc trưng; (ii) chọn mẫu trong công tác nghiên cứu khoa học nói chung, và cụ thể đối với phân tích thống kê và khai phá dữ liệu; (iii) xử lý dữ liệu trực tuyến, gồm các công cụ hiển thị, xử lý dữ liệu, sử dụng trong phân tích dữ liệu và trợ giúp quyết định; và (iv) sai sót hay mắc trong quá trình phân tích thống kê và khai phá dữ liệu, giúp người dùng tránh các lỗi thường gặp.

Một số kĩ thuật phân tích dữ liệu

Về loại dữ liệu cần phân tích thống kê hay khai phá dữ liệu, người ta đề cập (i) dữ liệu số; (ii) dữ liệu phạm trù; (iii) dữ liệu đồ thị; (iv) dữ liệu nhân... Trong chương này, ba loại dữ liệu chính được xem xét. Các kĩ thuật phân tích thống kê được trình bày theo các loại dữ liệu này. Một số khái niệm về thống kê được chính xác hóa qua các định nghĩa trong chương.

4.1. Các thuộc tính số

Phần này đề cập các phương pháp thống kê cơ bản để khai phá các thuộc tính số. Một số khía cạnh xét đến là (i) độ đo tán xạ; (ii) độ đo phụ thuộc tuyến tính; (iii) kết hợp giữa các thuộc tính. Cần tập trung vào kết nối giữa xác suất và hình học, đại số của ma trận dữ liệu.

4.1.1. Phân tích một biến

4.1.1.1. Giới thiệu về phân tích một biến

Phân tích một biến tập trung vào thuộc tính đơn, tại một thời điểm; nên ma trận dữ liệu D có thể xem như ma trận $n \times 1$, hay đơn giản là vectơ cột

$$D = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

trong đó X là thuộc tính số, $x_i \in R$. X được xem là biến ngẫu nhiên; mỗi điểm x_i như biến ngẫu nhiên nhận biết được. Mỗi biến x_i là độc lập và được phân phối đều như X . X được xem như vector n chiều, viết là $X \in R^n$.

Nói chung, mật độ xác suất hay hàm mật độ¹ $f(x)$ và hàm phân phối tích lũy² $F(x)$ đối với thuộc tính X đều là chưa biết. Tuy nhiên người ta có thể ước lượng các phân phối này trực tiếp từ mẫu dữ liệu, và nó cũng cho phép tính thống kê để ước lượng vài tham số quan trọng khác.

Hàm phân phối tích lũy thực nghiệm: I là biến nhị phân, cho biết sự thỏa mãn của điều kiện. Để có CDF thực nghiệm, cần tính với mọi $x \in R$, có bao nhiêu điểm nhỏ hơn hay bằng x . CDF thực nghiệm $\hat{F}$ cho biết nó được ước lượng cho các hàm phân phối tích lũy F .

Định nghĩa: Hàm phân phối tích lũy thực nghiệm CDF của X là
$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x); \text{ trong đó } I(x_i \leq x) = 1 \text{ nếu } x_i \leq x; 0 \text{ nếu ngược lại.}$$

Hàm phân phối tích lũy ngược: CDF ngược cho biết giá trị nhỏ nhất của X , có q phần các giá trị là nhỏ hơn, và $1 - q$ phần của các giá trị là lớn hơn. Người ta tính hàm phân phối tích lũy ngược thực nghiệm theo định nghĩa trên.

Định nghĩa: Hàm phân phối tích lũy ngược, hay hàm phân vị³, đối với biến ngẫu nhiên X là $F^{-1}(q) = \min\{x \mid F(x) \geq q\}$, $q \in [0, 1]$.

Hàm mật độ xác suất thực nghiệm: hàm PMF thực nghiệm cũng lấy mật độ xác suất $1/n$ tại mỗi điểm x_i .

Định nghĩa: Hàm mật độ xác suất thực nghiệm PMF¹ của X là
$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x); \text{ trong đó } I(x_i = x) = 1 \text{ nếu } x_i = x; \text{ ngược lại } 0.$$

¹ Mass function: hàm mật độ.

² CDF: Cumulative distribution function: hàm phân phối tích lũy.

³ Quantile function: hàm phân vị.

4.1.1.2. Các độ đo hướng tâm

Các độ đo này cho biết sự chụm của mật độ xác suất, các giá trị ở giữa.

Trung bình: Trung bình, hay giá trị kì vọng, của biến ngẫu nhiên X là trung bình số học của các giá trị của X . Nói ngắn gọn là con số về vị trí hay hướng tâm của phân phối của X .

Định nghĩa: Trung bình, hay giá trị kì vọng, của biến ngẫu nhiên rời rạc X được xác định $\mu = E[X] = \sum_x x * f(x)$; trong đó $f(x)$ là hàm mật độ xác suất của X .

Định nghĩa: Giá trị kì vọng của biến ngẫu nhiên liên tục X được xác định $\mu = E[X] = \int_{-\infty}^{\infty} x * f(x)dx$; trong đó $f(x)$ là hàm mật độ xác suất của X .

Trung bình mẫu: Nhận xét thấy $\hat{\mu} = \sum_x x * \hat{f}(x) = \sum_x x * (\frac{1}{n} \sum_{i=1}^n I(x_i = x)) = \frac{1}{n} \sum_{i=1}^n x_i$,

Định nghĩa: Trung bình mẫu² là hàm thống kê $\hat{\mu} = \{x_1, \dots, x_n\} \rightarrow R$, được xác định như trung bình của các x_i : $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$.

Trung bình mẫu không lệch: Ước lượng $\hat{\theta}$ được gọi là ước lượng không lệch đối với tham số θ nếu $E[\hat{\theta}] = \theta$ đối với mọi giá trị của θ . Trung bình mẫu $\hat{\mu}$ là ước lượng không lệch đối với trung bình dân cư μ , do $E[\hat{\mu}] = E[\frac{1}{n} \sum_{i=1}^n x_i] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$. Ở đây đã sử dụng việc các biến ngẫu nhiên x_i là IID đối với X , ngầm định rằng chúng có cùng trung bình μ như X , tức $E[x_i] = \mu$ với mọi x_i ; hàm kì vọng E là toán tử tuyến tính, tức đối với mọi biến ngẫu nhiên X, Y , số thực a, b , người ta có $E[ax + bY] = aE[X] + bE[Y]$.

¹ PMF: Probability mass function: hàm mật độ xác suất.

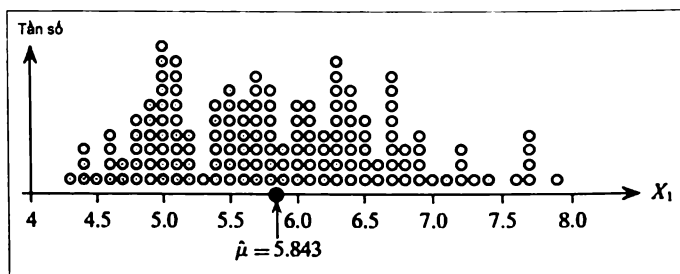
² Sample mean: trung bình mẫu.

Bền vững: Thống kê là *bền vững*¹ nếu nó không bị tác động bởi các giá trị cực đoan, chẳng hạn dữ liệu lẻ, trong dữ liệu. Trung bình mẫu không bền vững, do chỉ một giá trị lẻ lớn có thể làm lệch trung bình. Độ đo bền vững hơn là *trung bình tĩa*², thực hiện sau khi bỏ bớt phần nhỏ các dữ liệu cực đoan, về cả hai phía. Trung bình đôi khi gây hiểu lầm rằng nó không là giá trị xuất hiện trong mẫu, thậm chí không là giá trị của biến ngẫu nhiên.

Trung vị: *Trung vị mẫu* thu được theo CDF hay CDF ngược là $\hat{F}(m) = 0.5$ hay $m = \hat{F}^{-1}(0.5)$. Cách đơn giản để tính trung vị là (i) sắp xếp các giá trị x_i theo chiều tăng; (ii) nếu n lẻ, trung vị là giá trị tại vị trí $(n + 1) / 2$; nếu n chẵn, các giá trị tại vị trí $n/2$ hay $n/2 + 1$ đều là trung vị.

Định nghĩa: *Trung vị*³ của biến ngẫu nhiên là giá trị m , sao cho $P(X \leq m) \geq 1/2$ và $P(X \geq m) \leq 1/2$.

Không như trung bình, trung vị là bền vững, không chịu ảnh hưởng của các dữ liệu lẻ, và nó xuất hiện trong mẫu và là giá trị của biến ngẫu nhiên.



Hình 4.1. Trung bình mẫu đối với thuộc tính độ dài lá dài

Mode: *Mode*, hay *mẫu*, là giá trị mà hàm mật độ xác suất đạt giá trị cực đại, tức $\text{mode}(X) = \arg \max_x \hat{f}(x)$. Mode có thể không là độ đo hữu ích về hướng tâm đối với mẫu, do phân tử không đại diện có thể là phân tử phổ biến nhất.

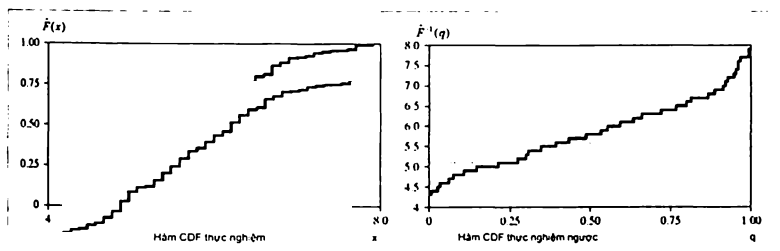
¹ Robust: bền vững.

² Trimmed mean: trung bình tĩa.

³ Median: trung vị.

Định nghĩa: Mode của biến ngẫu nhiên X là giá trị mà hàm mật độ xác suất đạt giá trị cực đại.

Về trung bình mẫu, trung vị và mode, xét thuộc tính độ dài lá đài (X_i) trong cơ sở dữ liệu hoa diên vĩ. Trung bình mẫu là $\bar{\mu} = (5.9 + 6.9 + \dots + 7.7 + 5.1) / 150 = 876.5 / 150 = 5.843$. Hình vẽ cho thấy 150 giá trị của độ dài lá đài và trung bình mẫu. Hình tiếp theo cho biết CDF thực nghiệm và CDF ngược thực nghiệm. Do $n = 150$ là chẵn, trung vị mẫu tại $n/2 = 75$ và $n/2 + 1 = 75$, trong thứ tự đã xếp. Từ CDF ngược, có thể thấy $\hat{F}(5.8) = 0.5$ hay $5.8 = \hat{F}^{-1}(0.5)$. Mode mẫu là 5, có thể quan sát thấy từ tần suất 5 trong hình. Mật độ xác suất tại $x = 5$ là $\hat{f}(5) = 10/150 = 0.067$.



Hình 4.2. CDF thực nghiệm và CDF ngược

4.1.1.3. Độ đo tán xạ

Các độ đo về tán xạ cho biết sự tán mát hay thay đổi về giá trị của biến ngẫu nhiên.

Khoảng: Khoảng của X là tham số lực lượng, khác với khoảng của hàm X , tức tập tất cả các giá trị mà X nhận. *Khoảng mẫu* là thống kê, được xác định $\hat{r} = \max_{i=1}^n \{x_i\} - \min_{i=1}^n \{x_i\}$. Khoảng nhạy cảm với các giá trị cực đoan, nên không bền vững.

Định nghĩa: Khoảng giá trị, hay dài, của biến ngẫu nhiên X là sai khác giữa các giá trị lớn nhất và nhỏ nhất của X , $r = \max \{X\} - \min \{X\}$.

Khoảng tứ phân vị: Điểm tứ phân¹ là giá trị đặc biệt của hàm phân vị, chia dữ liệu thành bốn phần. Các điểm phân vị là các giá

¹ Quartile: điểm tứ phân.

trị tứ phân 0.25, 0.5, 0.75 và 1.0. (i) Tứ phân thứ nhất là giá trị $q_1 = \hat{F}^{-1}(0.25)$, bên trái nó là 25% các điểm; (ii) tứ phân thứ hai có giá trị $q_2 = \hat{F}^{-1}(0.5)$, có 50% các điểm bên trái nó; (iii) tứ phân thứ ba $q_3 = \hat{F}^{-1}(0.75)$, có 75% các điểm thuộc bên trái nó; (iv) tứ phân thứ tư $q_4 = \hat{F}^{-1}(1.0)$ là giá trị lớn nhất của X , có 100% các điểm bên trái nó.

Độ đo bên vững hơn về mức tán xạ của X là *khoảng tứ phân vị* IQR^1 , được xác định $IQR = q_3 - q_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$. IQR cũng được xem như *khoảng xén*, trong đó người ta bỏ đi 25% các giá trị thấp và cao hơn của X . Nó là khoảng 50% giữa của các giá trị của X . IQR là bên vững. IQR mẫu được xác định theo CDF ngược thực nghiệm $I\hat{Q}R = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$.

Độ lệch chuẩn và độ lệch phương sai: Phương sai là giá trị kì vọng của độ lệch bình phương so với trung bình, được xác định theo $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$ khi X là biến rời rạc, và là $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$ khi biến X là liên tục.

Định nghĩa: Phương sai của biến ngẫu nhiên X đảm bảo độ đo về độ lệch của các giá trị của X so với trung bình hay giá trị kì vọng của X .

Độ lệch chuẩn σ là căn của phương sai σ^2 . Có thể viết phương sai như hiệu giữa kì vọng của X^2 và căn của kì vọng của X : $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - (E[X])^2$. Nhận xét thấy *moment thứ hai về trung bình*, ứng với $r = 2$, là trường hợp đặc biệt của *moment thứ r về trung bình* đối với biến ngẫu nhiên X , xác định là $E[(x - \mu)^r]$.

Phương sai mẫu: phương sai mẫu được xác định $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$.

Đó là độ lệch bình phương trung bình của các giá trị x_i so với trung bình mẫu $\hat{\mu}$, và có thể rút ra nhờ đặt hàm xác suất thực nghiệm $\hat{f}$ vào công thức định nghĩa σ^2 . $\hat{\sigma}^2 = \sum_x (x - \hat{\mu})^2 \hat{f}(x)$

¹ IQR Interquartile range: khoảng tứ phân vị.

$$= \sum_x (x - \bar{\mu})^2 \left(\frac{1}{n} \sum_{i=1}^n I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2. \text{ Độ lệch chuẩn mẫu}$$

là căn dương của phương sai mẫu: $\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2}$. Tỉ số chuẩn, hay tỉ số z^1 , của giá trị x_i là số các lệch chuẩn mà giá trị này cách trung bình $z_i = (x_i - \bar{\mu}) / \sigma^2$. Tỉ lệ này của x_i đo độ lệch của x_i so với giá trị trung bình $\bar{\mu}$, theo đơn vị đo của σ^2 .

Diễn giải hình học của phương sai mẫu: có thể xem mẫu dữ liệu đối với thuộc tính X như vectơ trong không gian n chiều, n là kích thước mẫu. Người ta viết $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$. Cho

$$Z = X - 1 * \bar{\mu} = \begin{pmatrix} x_1 - \bar{\mu} \\ \dots \\ x_n - \bar{\mu} \end{pmatrix}, \text{ kí hiệu vectơ thuộc tính trừ đi trung bình,}$$

trong đó $1 \in \mathbb{R}^n$ là vectơ n chiều với các phần tử 1. Có thể viết lại phương sai mẫu $\hat{\sigma}^2 = \frac{1}{n} \|Z\|^2 = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2$. Phương sai mẫu được diễn giải như độ lớn bình phương của vectơ thuộc tính giữa, hay tích chấm của vectơ thuộc tính giữa với nó, đã chuẩn hóa theo kích thước mẫu.

Tiếp tục thí dụ về độ dài lá đài hoa, thấy khoảng mẫu là $\max_i\{x_i\} - \min_i\{x_i\} = 7.9 - 4.3 = 3.6$. Theo CDF ngược, có thể phát hiện IQR mẫu theo: $\hat{q}_1 = \hat{F}^{-1}(0.25) = 5.1$; $\hat{q}_3 = \hat{F}^{-1}(0.75) = 6.4$; $IQR = \hat{q}_3 - \hat{q}_1 = 6.4 - 5.1 = 1.3$. Phương sai mẫu có thể tính theo vectơ dữ liệu giữa $\hat{\sigma}^2 = \frac{1}{n} (X - 1.\bar{\mu})^T (X - 1.\bar{\mu}) = 102.168/150 = 0.681$. Khi đó độ lệch chuẩn mẫu là $\hat{\sigma} = \sqrt{0.681} = 0.825$.

Phương sai của trung bình mẫu: do tự trung bình mẫu $\bar{\mu}$ là thống kê, người ta có thể tính giá trị trung bình và phương sai của nó. Giá trị kì vọng của trung bình mẫu đơn giản là μ . Để có

¹ Z - score: tỉ số z .

phương sai của trung bình mẫu, nhận xét thấy các biến ngẫu nhiên x_i là độc lập, nên $\text{var}(\sum_{i=1}^n x_i) = \sum_{i=1}^n \text{var}(x_i)$. Ngoài ra, do các x_i có phân phối như X , chúng có cùng phương sai như X , tức $\text{var}(x_i) = \sigma^2$ đối với mọi i . Kết hợp với nhau, $\text{var}(\sum_{i=1}^n x_i) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n \sigma^2 = n \cdot \sigma^2$. Ngoài ra, $E[\sum_{i=1}^n x_i] = n\mu$ nên $\text{var}(\bar{\mu}) = E[(\bar{\mu} - \mu)^2] = E[\bar{\mu}^2] - \mu^2 = E[(\frac{1}{n} \sum_{i=1}^n x_i)^2] - \frac{1}{n^2} E[\sum_{i=1}^n x_i]^2 =$
 $= \frac{1}{n^2} (E[(\sum_{i=1}^n x_i)^2] - E[\sum_{i=1}^n x_i]^2) = \frac{1}{n^2} \text{var}(\sum_{i=1}^n x_i) = \frac{\sigma^2}{n}$.

Nói cách khác, trung bình mẫu $\bar{\mu}$ thay đổi hay độ lệch so với trung bình μ tỉ lệ với phương sai lực lượng σ^2 . Tuy nhiên, độ lệch sẽ nhỏ đi nếu n lớn.

Thiên lệch của phương sai mẫu: Phương sai mẫu là thiên lệch, nhưng không thiên lệch theo cách tiệm cận. Phương sai mẫu là ước lượng thiên lệch đối với phương sai lực lượng thực σ^2 , tức $E[\hat{\sigma}^2] \neq \sigma^2$. Xét $\sum_{i=1}^n (x_i - \mu) = n \cdot (\bar{\mu} - \mu)^2 + \sum_{i=1}^n (x_i - \bar{\mu})^2$. Tính kì vọng $\hat{\sigma}^2$:
 $E[\hat{\sigma}^2] = E[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2] = E[\sum_{i=1}^n (x_i - \mu)^2] - E[(\bar{\mu} - \mu)^2]$. Lưu ý các biến ngẫu nhiên x_i là IID theo X , có nghĩa có cùng trung bình μ và phương sai σ^2 như X , tức $E[(x_i - \mu)^2] = \sigma^2$. Ngoài ra trung bình mẫu $\bar{\mu}$ có phương sai $E[(\bar{\mu} - \mu)^2] = \sigma^2 / n$. Thay thế trong biểu thức, được $E[\hat{\sigma}^2] = n\sigma^2 / n - \sigma^2 / n = ((n - 1) / n) \cdot \sigma^2$. Phương sai mẫu $\hat{\sigma}^2$ là ước lượng lệch của σ^2 , do giá trị kì vọng của nó khác với phương sai lực lượng với hệ số $(n - 1) / n$. Tuy nhiên nó là không thiên lệch tiệm cận, tức sự thiên lệch mất đi khi $n \rightarrow \infty$. Tức $E[\hat{\sigma}^2] \rightarrow \sigma^2$ khi $n \rightarrow \infty$.

4.1.2. Phân tích hai biến

Phân tích hai biến làm việc cùng lúc với hai thuộc tính đối tượng. Hai thuộc tính dữ liệu số đề cập ở đây là X_1, X_2 , thể hiện như dữ liệu D theo ma trận $n \times 2$:

$$D = \begin{pmatrix} X_1 & X_2 \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Về hình học, có thể xem (i) D như n điểm hay n vector trong không gian hai chiều, trên thuộc tính X_1, X_2 , tức $x_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$; (ii) D như hai điểm hai vector trong không gian n chiều, gồm các điểm, tức mỗi cột là một vector trong $\mathbb{R}^n$, $X_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$.

Theo xác suất, vector cột $X = (X_1, X_2)^T$ được xem như biến ngẫu nhiên vector hai biến, và điểm $x_i, i = 1..n$, được xem như mẫu ngẫu nhiên lấy từ X , tức x_i được xét độc lập và được phân phối như X .

Định nghĩa: hàm mật độ xác suất kết hợp thực nghiệm đối với X được xác định $\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x)$; $\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^n I(x_{i1} = x_1, x_{i2} = x_2)$.

Trong định nghĩa trên, $x = (x_1, x_2)^T$ và I là biến có giá trị 1 khi đối số đúng, 0 nếu ngược lại, tức $I(x_i = x) = 1$ nếu $x_{i1} = x_1$ và $x_{i2} = x_2$. Như trong trường hợp phân tích một biến, hàm xác suất đạt mật độ xác suất $1/n$ tại mỗi điểm trong mẫu dữ liệu.

4.1.2.1. Các đo về vị trí và tán xạ

Trung bình: Vector trung bình hai biến đơn giản là vector của các giá trị kì vọng theo mỗi thuộc tính.

Định nghĩa: Trung bình hai biến là giá trị kì vọng của biến ngẫu nhiên vector X , được xác định $\mu = E[X] = E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$.

Trung bình mẫu tính theo hàm mật độ xác suất thực nghiệm $\hat{f}_{X_1}$ và $\hat{f}_{X_2}$ của X_1, X_2 . Nó cũng được tính như PMF kết hợp thực nghiệm: $\hat{\mu} = \sum_x x * \hat{f}(x) = \sum_x x * (\frac{1}{n} \sum_{i=1}^n I(x_i = x)) = \frac{1}{n} \sum_{i=1}^n x_i$.

Phương sai: Tính phương sai theo mỗi thuộc tính, được σ^2_1 và σ^2_2 , rồi tính *tổng phương sai* var (D) = $\sigma^2_1 + \sigma^2_2$. Các phương sai mẫu $\hat{\sigma}^2_1$ và $\hat{\sigma}^2_2$ tính được theo công thức một biến, rồi tính *tổng phương sai mẫu*, $\hat{\sigma}^2_1 + \hat{\sigma}^2_2$.

4.1.2.2. Các độ đo kết hợp

Hiệp phương sai: Do tính tuyến tính của kì vọng, người ta có:

$$\begin{aligned}\sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - E[X_1]\mu_2 - E[X_2]\mu_1 + \mu_1\mu_2 \\ &= E[X_1X_2] + \mu_1\mu_2 = E[X_1X_2] + E[X_1]E[X_2]\end{aligned}$$

Định nghĩa: Hiệp phương sai giữa hai thuật toán X_1 và X_2 là độ đo về kết hợp, hay phụ thuộc tuyến tính giữa chúng, được xác định: $\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$.

Nó cho thấy khái quát về phương sai hai biến. Nếu X_1, X_2 là các biến ngẫu nhiên độc lập, thì $E[X_1X_2] = E[X_1]E[X_2]$, nên hiệp phương sai bằng 0. Tuy nhiên, điều ngược lại không đúng, tức khi $\sigma_{12} = 0$, không rút ra X_1, X_2 độc lập, mà chỉ thấy không có phụ thuộc tuyến tính giữa chúng. Tức không thể có quan hệ thứ tự cao hay sự phụ thuộc giữa hai thuộc tính.

Định nghĩa: Hiệp phương sai mẫu giữa X_1 và X_2 là $\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$.

Nó có thể thu được nhờ thay thế trong $\hat{f}(x_1, x_2)$:

$$\hat{\sigma}_{12} = E[(X_1 - \hat{\mu}_1)(X_2 - \hat{\mu}_2)] = \sum_{x=(x_1, x_2)^T} (x_1 - \hat{\mu}_1)(x_2 - \hat{\mu}_2) \hat{f}(x_1, x_2)$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{x=(x_1, x_2)^T} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2) * I(x_{i1} = x_1, x_{i2} = x_2) = \\
 &= \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)
 \end{aligned}$$

Hiệp phương sai mẫu là khái quát của phương sai mẫu, do $\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 = \hat{\sigma}_1^2$.

Tương quan: có định nghĩa như sau.

Định nghĩa: Tương quan giữa các biến X_1, X_2 là hiệp phương sai chuẩn hóa, nhờ chuẩn hóa hiệp phương sai với độ lệch chuẩn của mỗi

biến $\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$.

Định nghĩa về tương quan mẫu.

Định nghĩa: Tương quan mẫu đối với X_1, X_2 là

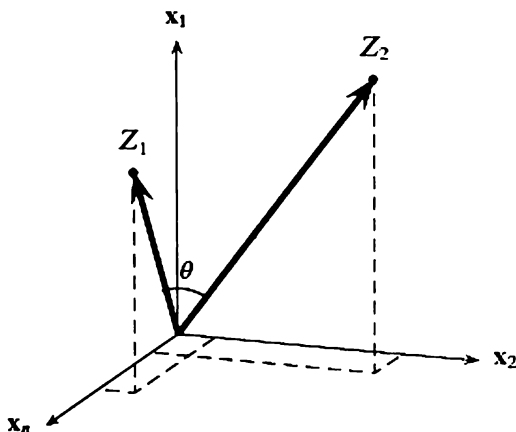
$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}.$$

Diễn giải hình học của hiệp phương sai và tương quan mẫu: kí hiệu Z_1, Z_2 là các vector thuộc tính hướng tâm¹ trong R^n :

$$Z_1 = X_1 - 1 \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \dots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix}; \quad Z_2 = X_2 - 1 \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \dots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}.$$

Hiệp phương sai mẫu được như $\hat{\sigma}_{12} = Z_1^T Z_2 / n$. Nói khác đi, hiệp phương sai giữa hai thuộc tính là tích chấm giữa hai vector thuộc tính hướng tâm, được chuẩn hóa theo kích thước mẫu. Đó là khái quát của phương sai mẫu một biến.

¹ Centered: hướng tâm.



Hình 4.3. Diễn giải hình học về hiệp phương sai và tương quan

Trong hình diễn giải, hai vectơ thuộc tính hướng tâm là trong không gian n chiều R^n , có n điểm.

Tương quan mẫu có thể viết lại, như cosin của góc giữa hai vectơ thuộc tính hướng tâm.

$$\hat{\rho}_{12} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1} \sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\| \|Z_2\|} = \left(\frac{Z_1}{\|Z_1\|} \right)^T \left(\frac{Z_2}{\|Z_2\|} \right) = \cos \theta.$$

Ma trận hiệp phương sai: Thông tin về hiệp phương sai, phương sai đối với hai thuộc tính X_1, X_2 được thể hiện trong *ma trận hiệp phương sai* 2×2 :

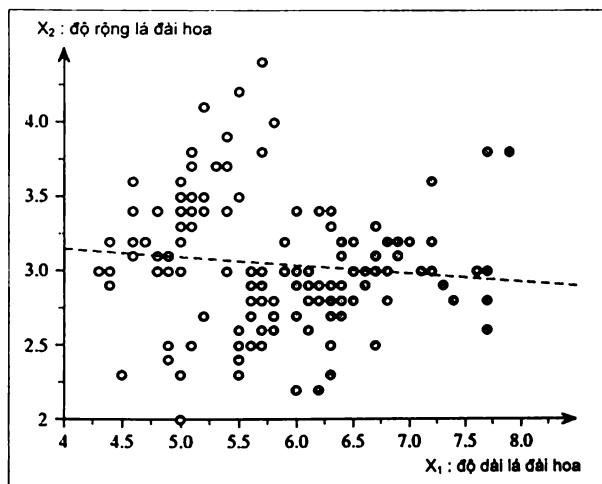
$$\begin{aligned} \Sigma &= E[(X - \mu)(X - \mu)^T] = E \left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \end{aligned}$$

Do $\sigma_{12} = \sigma_{21}$, Σ là ma trận đối xứng. Ma trận hiệp phương sai cho biết các phương sai về thuộc tính trên đường chéo chính, và thông tin hiệp phương sai trên các phần tử ngoài đường chéo chính.

Định nghĩa: Phương sai tổng cộng¹ của hai thuộc tính là tổng của các phần tử trên đường chéo chính của Σ , gọi là vết của Σ , là $\text{var}(D) = \text{tr}(\Sigma) = \sigma_1^2 + \sigma_2^2$.

Phương sai tổng quát² của hai thuộc tính cũng xem như hiệp phương sai, bổ sung vào các phương sai, tức theo định nghĩa về tương quan $\sigma_{12} = \rho_{12} \sigma_1 \sigma_2$; người ta có định thức của ma trận hiệp phương sai Σ , kí hiệu $|\Sigma|$ hay $\det(\Sigma)$. $|\Sigma| = \det(\Sigma) = \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 = \sigma_1^2 \sigma_2^2 - \rho_{12}^2 \sigma_1^2 \sigma_2^2 = (1 - \rho_{12}^2) \sigma_1^2 \sigma_2^2$. Do $|\rho_{12}| \leq 1$, nên $\det(\Sigma) \geq 0$.

Ma trận hiệp phương sai mẫu là $\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 \end{pmatrix}$, có cùng tính chất như Σ , tức đối xứng và không âm.



Hình 4.4. Hiệp phương sai giữa độ dài và độ rộng lá đài hoa

Về trung bình mẫu và hiệp phương sai, xét thí dụ sử dụng dữ liệu về độ dài và độ rộng lá đài hoa, các điểm $n = 150$, không gian hai chiều.

¹ Total variance: phương sai tổng cộng.

² Generalized variance: phương sai tổng quát.

Trung bình mẫu là $\hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$; ma trận hiệp phương sai là $\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$. Phương sai đối với độ dài lá đài hoa là $\hat{\sigma}_1^2 = 0.681$ và đối với độ rộng lá đài hoa là $\hat{\sigma}_2^2 = 0.187$. Hiệp phương sai giữa hai thuộc tính là $\hat{\sigma}_{12} = -0.039$ và tương quan giữa chúng là $\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$. Do vậy, tương quan là âm, thấp, với việc có đường phù hợp tuyến tính trong hình.

Người ta có thể xem hai thuộc tính như hai điểm trong $\mathbb{R}^n$. Tương quan lúc này là cosin của góc giữa chúng, tức $\hat{\rho}_{12} = \cos\theta = -0.109$, cho biết $\theta = 96.26^\circ$. Góc này gần với 90° , có nghĩa hai vectơ hầu như trực giao, tức tương quan yếu. Ngoài ra, góc này lớn hơn 90° , cho biết tương quan âm.

Phương sai mẫu tổng cộng là $tr(\hat{\Sigma}) = 0.681 + 0.187 = 0.868$, và phương sai tổng quát mẫu $|\hat{\Sigma}| = \det(\hat{\Sigma}) = 0.681 \cdot 0.187 - (-0.039)^2 = 0.126$.

4.1.3. Phân tích đa biến

Phân tích đa biến xét tất cả d thuộc tính $X_1, \dots, X_d$. Dữ liệu là ma trận $n \times d$:

$$D = \begin{pmatrix} X_1 & X_2 & \dots & X_d \\ x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}$$

Theo dòng, dữ liệu này được xem như tập của n điểm hay vectơ trong không gian d thuộc tính $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T \in \mathbb{R}^d$. Nhìn theo cột, dữ liệu được xem như tập d điểm hay vectơ trong không gian n chiều, có các điểm dữ liệu $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T \in \mathbb{R}^n$. Theo quan điểm xác suất, d thuộc tính được mô hình hóa như biến ngẫu

nhien vector, $X = (X_1, X_2, \dots, X_d)^T$, và điểm x_i được xem như mẫu ngẫu nhiên lấy từ X , tức chúng là độc lập và có phân phối như X .

4.1.3.1. Trung bình

Định nghĩa: Vector trung bình đa biến là trung bình của mỗi thuộc tính

$$\mu = E[X] = E\left[\begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \dots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_d \end{pmatrix}$$

Tổng quát hóa các định nghĩa trước, trung bình mẫu là

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

4.1.3.2. Ma trận hiệp phương sai

Tổng quát từ 2 chiều cho d chiều, ma trận hiệp phương sai là ma trận đối xứng $d \times d$, cho biết phương sai giữa mỗi cặp thuộc tính.

$$\Sigma = E[(X - \mu)(X - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \sigma_{2d} \\ \dots & \dots & \dots \\ \sigma_{d1} & \sigma_{d2} & \sigma_d^2 \end{pmatrix}$$

Phần tử chéo σ_i^2 xác định phương sai thuộc tính đối với X_i , còn các phần tử khác thể hiện hiệp phương sai giữa cặp X_i, X_j .

4.1.3.3. Ma trận hiệp phương sai

Ma trận hiệp phương sai trong trường hợp đa biến là nửa xác định dương¹. Tức $a^T \Sigma a \geq 0$ đối với vectơ d chiều a .

Có $a^T \Sigma a = a^T E[(X - \mu)(X - \mu)^T] a = E[a^T (X - \mu)(X - \mu)^T a] = E[Y^2] \geq 0$; trong đó $Y = a^T (X - \mu) = \sum_{i=1}^n a_i (X_i - \mu_i)$.

¹ Semidefinite: nửa xác định.

Do Σ là đối xứng, ngầm định rằng các giá trị riêng¹ của Σ là thực và không âm. Nói khác đi, d giá trị riêng của Σ có thể bố trí từ lớn nhất đến nhỏ nhất: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Định thức của Σ là $\det(\Sigma) = \prod_{i=1}^d \lambda_i \geq 0$.

4.1.3.4. Phương sai tổng cộng và phương sai tổng quát

Định nghĩa: Phương sai tổng cộng là vết của ma trận hiệp phương sai, $\text{var}(D) = \text{tr}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_d^2$.

Định nghĩa: Phương sai tổng quát là định thức của ma trận hiệp phương sai, $\det(\Sigma)$.

4.1.3.5. Ma trận hiệp phương sai mẫu

Ma trận hiệp phương sai mẫu là $\hat{\Sigma} = E[(X - \hat{\mu})(X - \hat{\mu})^T]$

$$= \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots & \hat{\sigma}_{1d} \\ \dots & & & \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \dots & \hat{\sigma}_d^2 \end{pmatrix}.$$

Thay vì tính từng phần tử ma trận hiệp

phương sai, người ta sử dụng các phép ma trận. Lấy Z thể hiện ma trận dữ liệu hướng tâm, tức ma trận của vector thuộc tính hướng tâm $Z_i = X_i - 1 \cdot \hat{\mu}_i$, trong đó $1 \in \mathbb{R}^n$. $Z = D - 1 \cdot \hat{\mu}_i = (Z_1 \ Z_2 \dots Z_d)$. Hoặc ma trận dữ liệu hướng tâm được viết theo các điểm hướng tâm $z_i = x_i - \hat{\mu}$; và cả ma trận hiệp phương sai mẫu:

$$Z = D - 1 \cdot \hat{\mu}^T = \begin{pmatrix} x_1^T - \hat{\mu}^T \\ x_2^T - \hat{\mu}^T \\ \dots \\ x_n^T - \hat{\mu}^T \end{pmatrix} = \begin{pmatrix} z_1^T \\ z_2^T \\ \dots \\ z_n^T \end{pmatrix};$$

$$\hat{\Sigma} = \frac{1}{n}(Z^T Z) = \frac{1}{n} \begin{pmatrix} Z_1^T Z_1 & Z_1^T Z_2 & \dots & Z_1^T Z_d \\ Z_2^T Z_1 & Z_2^T Z_2 & \dots & Z_2^T Z_d \\ \dots & & & \\ Z_d^T Z_1 & Z_d^T Z_2 & \dots & Z_d^T Z_d \end{pmatrix}$$

¹ Eigenvalue: giá trị riêng.

Ma trận hiệp phương sai thể hiện như cặp *tích chấm trong* của các vector thuộc tính hướng tâm, được chuẩn hóa theo kích thước mẫu. Nhìn theo các điểm z_i , ma trận hiệp phương sai cũng được viết như tổng các ma trận hạng một, như *tích chấm ngoài*¹ của mỗi điểm hướng tâm: $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n z_i \cdot z_i^T$.

Về trung bình mẫu và ma trận hiệp phương sai, xét bốn thuộc tính số, tên là *độ dài lá đài hoa*, *độ rộng lá đài*, *độ dài cánh*, *độ rộng cánh*. Vector trung bình mẫu nhiều biến là $\hat{\mu} = (5.843 \ 3.054 \ 3.759 \ 1.199)^T$ và ma trận hiệp phương sai mẫu là:

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 & 1.265 & 0.513 \\ -0.039 & 0.187 & -0.320 & -0.117 \\ 1.265 & 0.320 & 3.092 & 1.288 \\ 0.513 & -0.117 & 1.288 & 0.579 \end{pmatrix}$$

Phương sai tổng cộng mẫu $\text{var}(D) = \text{tr}(\hat{\Sigma}) = 0.681 + 0.187 + 3.092 + 0.579 = 4.539$; và phương sai tổng quát là $\det(\hat{\Sigma}) = 1.87 \cdot 10^{-3}$.

Về tích chấm trong và ngoài, xét bộ dữ liệu hai chiều $D = \begin{pmatrix} A_1 & A_2 \\ 1 & 0.8 \\ 2 & 2.4 \\ 9 & 5.5 \end{pmatrix}$. Các vector trung bình là $\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} 15/3 \\ 8.7/3 \end{pmatrix} = \begin{pmatrix} 5 \\ 2.9 \end{pmatrix}$

và ma trận dữ liệu hướng tâm

$$Z = D - 1 \cdot \hat{\mu}^T = \begin{pmatrix} 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 5 & 2.9 \end{pmatrix} = \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

Theo tích chấm trong, tính ma trận hiệp phương sai mẫu được

$$\hat{\Sigma} = \frac{1}{n} Z^T Z = \frac{1}{3} \begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

¹ Inner dot product, outer dot product: tích chấm trong, tích chấm ngoài.

$$= \frac{1}{3} \begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}.$$

Tích chấm ngoài được:

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n z_i \cdot z_i^T \\ &= \frac{1}{3} \left(\begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot (-4 \ -2.1) + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot (0 \ -0.5) + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot (4 \ 2.6) \right) \\ &= \frac{1}{3} \left(\begin{pmatrix} 16.0 & 8.4 \\ 8.4 & 4.41 \end{pmatrix} + \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{pmatrix} + \begin{pmatrix} 16.0 & 10.4 \\ 10.4 & 6.76 \end{pmatrix} \right) = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}. \end{aligned}$$

4.1.4. Chuẩn hóa dữ liệu

Khi phân tích hai hay nhiều thuộc tính, thường cần chuẩn hóa các giá trị của các thuộc tính, đặc biệt khi các giá trị này khác nhau về thang đo.

4.1.4.1. Chuẩn hóa khoảng

Cho X là thuộc tính và $x_1, \dots, x_n$ là các mẫu ngẫu nhiên rút ra từ X . Việc *chuẩn hóa khoảng* cho phép mỗi giá trị lấy thang đo theo khoảng mẫu $\hat{r}$ của X , tức $x_i' = \frac{x_i - \min_i \{x_i\}}{\hat{r}} = \frac{x_i - \min_i \{x_i\}}{\max_i \{x_i\} - \min_i \{x_i\}}$.

Sau quá trình chuyển hóa, các giá trị thuộc tính thuộc vào $[0; 1]$.

4.1.4.2. Chuẩn hóa tỉ lệ chuẩn

Định nghĩa: Chuẩn hóa tỉ lệ chuẩn, hay chuẩn hóa z, thay mỗi giá trị thuộc tính bằng tỉ lệ z, tức $x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$.

Trong định nghĩa, $\hat{\mu}$ là trung bình mẫu, $\hat{\sigma}^2$ là phương sai mẫu của X . Sau chuyển hóa, giá trị mới có $\hat{\mu}' = 0$, và độ lệch chuẩn $\hat{\sigma}' = 1$.

Xét thuộc tính *tuổi* và *thu nhập* trong bảng. Chúng có thang đo khác nhau, và *thu nhập* có các giá trị lớn hơn nhiều. Xét khoảng cách giữa x_1, x_2 : $\|x_1 - x_2\| = \|(2, 200)^T\| = \sqrt{2^2 + 200^2} = 200.01$. Theo quan sát, tham gia của *tuổi* bị giá trị của *thu nhập* lấn át. Khoảng mẫu của *tuổi* là $\bar{r} = 40 - 12 = 28$, với giá trị nhỏ nhất là 12. Sau khi chuẩn hóa khoảng, giá trị mới là: $\text{tuổi}' = (0, 0.071, 0.214, 0.393, 0.536, 0.571, 0.786, 0.893, 0.964, 1)^T$. Chẳng hạn điểm $x_2 = (x_{21}, x_{22}) = (14, 500)$, giá trị $x_{21} = 14$ được chuyển thành $x_{21}' = (14 - 12)/28 = 0.071$.

Bảng 4.1. Dữ liệu cần chuẩn hóa

x_i	Tuổi (X_1)	Thu nhập (X_2)
x_1	12	300
x_2	14	500
x_3	18	1000
x_4	23	2000
x_5	27	3500
x_6	28	4000
x_7	34	4300
x_8	37	6000
x_9	39	2500
x_{10}	40	2700

Ngược lại, khoảng mẫu của *thu nhập* là $6000 - 300 = 5700$, với giá trị nhỏ nhất là 300. Nó được chuyển thành $\text{thu nhập}' = (0, 0.035, 0.123, 0.298, 0.561, 0.649, 0.702, 1, 0.386, 0.421)^T$, mà $x_{22} = 0.035$. Khoảng cách giữa x_1 và x_2 sau khi chuẩn hóa khoảng là $\|x_1' - x_2'\| = \|(0, 0)^T - (0.071, 0.035)^T\| = \|(-0.071, -0.035)^T\| = 0.079$. Lúc này *thu nhập* không làm lệch khoảng cách.

Đối với chuẩn z , cần tính trung bình và độ lệch chuẩn của cả hai thuộc tính. Với *tuổi*, $\bar{\mu} = 27.2$, $\bar{\sigma} = 9.77$; với *thu nhập*, $\bar{\mu} = 2680$, $\bar{\sigma} = 1726.15$. Thuộc tính *tuổi* được chuyển $\text{tuổi}' = (-1.56, -1.35, -0.94,$

$-0.43, -0.02, 0.08, 0.07, 1.0, 1.21, 1.31)^T$. Giá trị $x_{21} = 14$ đối với điểm $x_2 = (x_{21}, x_{22}) = (14, 500)$ được chuyển thành $x_{21}' = (14 - 27.2)/9.77 = -1.35$. Tương tự, thu nhập' $= (-1.38, -1.26, -0.97, -0.39, 0.48, 0.77, 0.94, 1.92, -0.10, 0.01)^T$. Vậy $x_{22} = -1.26$. Khoảng cách giữa x_1 và x_2 sau khi dùng chuẩn z là $\|x_1' - x_2'\| = \|(-1.56, -1.38)^T - (1.35, -1.26)^T\| = \|(-0.18, -0.12)^T\| = 0.216$.

4.1.5. Phân phối chuẩn

Phân phối chuẩn là một trong những hàm phân phối mật độ xác suất quan trọng nhất, đặc biệt do nhiều biến vật lý quan sát được tuân theo phân phối xấp xỉ chuẩn. Ngoài ra, phân phối lấy mẫu của trung bình của bất kì phân phối xác suất tùy ý tuân theo phân phối chuẩn. Phân phối chuẩn giữ vai trò như phân phối tham số lựa chọn trong phân cụm, ước lượng mật độ và phân lớp.

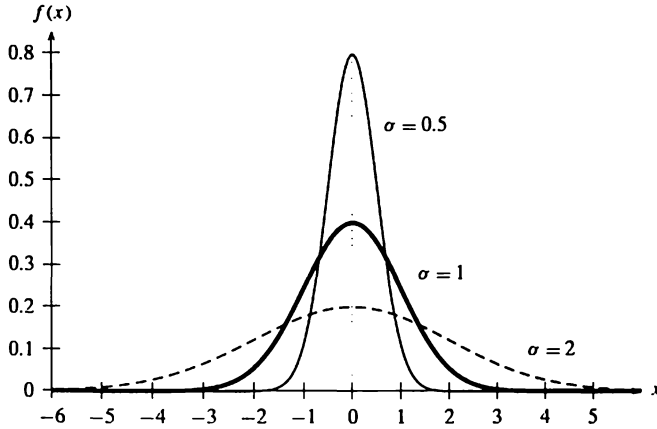
4.1.5.1. Phân phối chuẩn một biến

Định nghĩa: Biến ngẫu nhiên X có phân phối chuẩn, với tham số trung bình μ và phương sai σ^2 , nếu hàm phân phối xác suất của X là

$$f(x|\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}.$$

Hạng thức $(x - \mu)^2$ đo khoảng cách của biến x so với trung bình μ của phân phối, nên mật độ xác suất giảm theo hàm mũ như hàm của khoảng cách so với trung bình. Giá trị cực đại của mật độ xảy ra tại $x = \mu$, và $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}$, tỉ lệ nghịch với độ lệch chuẩn σ của phân phối.

Hình vẽ phân phối chuẩn, có $\mu = 0$, và $\sigma^2 = 1$. Phân phối chuẩn có hình *chuông*, đối xứng quanh trung bình. Đồ thị cho thấy hiệu ứng của giá trị sai khác của độ lệch chuẩn lên hình dạng của phân phối. Giá trị càng nhỏ, $\sigma = 0.5$ cho phân phối nhọn, phân rõ nhanh; giá trị càng lớn, $\sigma = 2$ cho phân phối mềm mòng, chậm phân rõ hơn.



Hình 4.5. Phân phối chuẩn với $\mu = 0$ và các phương sai khác nhau

Mật độ xác suất:

Định nghĩa: Mật độ xác suất, trong $[a, b]$, của phân phối chuẩn trong khoảng đó là $P(a \leq x \leq b) = \int_a^b f(x | \mu, \sigma^2) dx$.

Người ta đặc biệt quan tâm đến mật độ xác suất tập trung tại k lệch chuẩn từ trung bình, tức đối với khoảng $[\mu - k\sigma, \mu + k\sigma]$, tính

theo: $P(\mu - k\sigma \leq x \leq \mu + k\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu - k\sigma}^{\mu + k\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$. Thông qua

thay đổi của tham số $z = (x - \mu) / \sigma$, người ta có dạng phân phối chuẩn

$$\text{tức: } P(-k \leq z \leq k) = \frac{1}{\sqrt{2\pi}} \int_{-k}^k e^{-\frac{1}{2}z^2} dz = \frac{2}{\sqrt{2\pi}} \int_0^k e^{-\frac{1}{2}z^2} dz.$$

Do $e^{-\frac{1}{2}z^2}$ đối xứng, và tích phân trong $[-k, k]$ gấp đôi tích phân trong $[0, k]$, nên đổi biến $t = z / \sqrt{2}$, thu được:

$$P(-k \leq z \leq k) = 2P(0 \leq t \leq k / \sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_0^{k/\sqrt{2}} e^{-t^2} dt = \text{erf}(k / \sqrt{2}).$$

Trong đó $\text{erf}(k / \sqrt{2})$ là hàm sai số Gauss², được xác định như

¹ Standard normal distribution: phân phối chuẩn tắc, phân phối với $\mu = 0, \sigma = 1$.

² Erf error function: hàm sai sót.

$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. Khi $k = 1$, người ta có $P(\mu - \sigma \leq x \leq \mu + \sigma) = erf(1/\sqrt{2}) = 0.6827$, có nghĩa 68.27% các điểm nằm trong 1 lệch chuẩn so với trung bình. Với $k = 2$, có $erf(2/\sqrt{2}) = 0.9545$, và với $k = 3$, có $erf(3/\sqrt{2}) = 0.9973$. Vậy, hầu hết mật độ xác suất, tức 99.73%, của phân phối chuẩn nằm vào $\pm 3\sigma$ so với trung bình μ .

4.1.5.2. Phân phối chuẩn đa biến

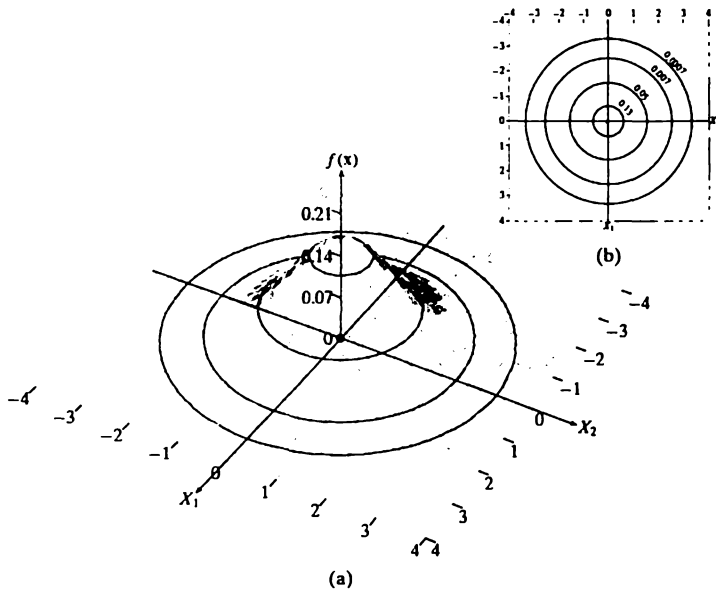
Cho biến X vectơ d chiều $X = (X_1, \dots, X_d)^T$; $|\Sigma|$ là định thức của ma trận hiệp phương sai.

Định nghĩa: X có phân phối chuẩn đa chiều, với trung bình tham số μ và ma trận hiệp phương sai Σ , nếu hàm mật độ xác suất nhiều biến kê hợp của nó được xác định $f(x|\mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} e(-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2})$.

Hạng thức $(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$ đo khoảng cách, gọi là *khoảng cách Mahalanobis*, của điểm x so với trung bình μ của phân phối, tính đến tất cả thông tin phương sai, hiệp phương sai giữa các thuộc tính. Khoảng cách Mahalanobis là tổng quát của khoảng cách Oclit, do khi lấy $\Sigma = I$, trong đó I là ma trận đơn vị $d \times d$, với đường chéo là 1, ngược lại là 0, người ta có: $(x_i - \mu)^T I^{-1} (x_i - \mu) = \|x_i - \mu\|^2$. Khoảng cách Oclit bỏ qua thông tin hiệp phương sai giữa các thuộc tính, trong khi khoảng cách Mahalanobis hiện tường mình.

Phân phối chuẩn tắc đa biến có tham số $\mu = 0$, $\Sigma = I$. Hình vẽ cho thấy mật độ xác suất của phân phối chuẩn tắc hai biến, $d = 2$, với tham số $\mu = 0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ và $\Sigma = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Điều này tương ứng với trường hợp hai biến độc lập, và cả hai theo phân phối chuẩn. Bản chất đối xứng của phân phối chuẩn tắc thấy rõ với hình đường viền. Mỗi mức của đường cong thể hiện tập các điểm x với mật độ cố định $f(x)$.



Hình 4.6. Mật độ chuẩn tắc hai biến và đường viền của nó, với $\mu = 0$ và $\Sigma = 1$

Hình học của chuẩn đa chiều: về khía cạnh hình học của phân phối chuẩn đa chiều đối với trung bình μ tùy ý và ma trận hiệp phương sai Σ , người ta so sánh với phân phối chuẩn tắc, có thể mong rằng các đường mật độ có hướng thay đổi, có tỉ lệ và xoay. Việc thay đổi và chuyển xuất phát từ việc trung bình μ không nhất thiết là 0. Tỉ lệ hay độ nghiêng là do phương sai của thuộc tính, và việc xoay là do hiệp phương sai.

Hình dạng, hay hình học, sẽ rõ hơn nếu xét phân rã vector riêng của ma trận hiệp phương sai; Σ là ma trận nửa xác định dương $d \times d$. Phương trình vector riêng đối với Σ là $\Sigma u_i = \lambda_i u_i$; trong đó λ_i là giá trị riêng của Σ và vectơ $u_i \in \mathbb{R}^d$ là vectơ riêng, ứng với λ_i . Do Σ đối xứng và nửa xác định dương, nó có d giá trị riêng thực, không âm, được sắp xếp theo thứ tự từ lớn nhất đến nhỏ nhất, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Ma trận đường chéo Λ được dùng để ghi các giá trị riêng:

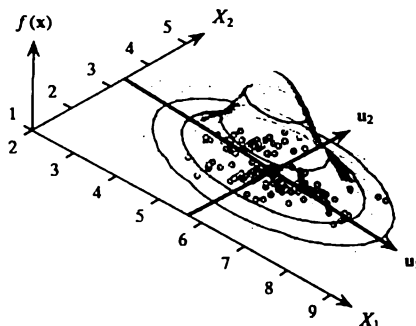
$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \lambda_d \end{pmatrix}.$$

Ngoài ra, các vectơ riêng là vectơ đơn vị, chuẩn, và trực giao lẫn nhau, tức $u_i^T u_i = 1$ đối với mọi i , $u_i^T u_j = 0$. Các vectơ riêng có thể đặt cùng nhau vào ma trận U , xác định như ma trận chuẩn và các cột trực giao $U = (u_1 \ u_2 \dots \ u_d)$. Phân rã vectơ riêng của Σ có thể được biểu diễn $\Sigma = U\Lambda U^T$. Đẳng thức này diễn giải về hình học như thay đổi theo các vectơ cơ sở. Từ d chiều ứng với d thuộc tính X_i , người ta rút ra d chiều mới u_i . Σ là ma trận hiệp phương sai trong không gian ban đầu, trong khi Λ là ma trận hiệp phương sai trong không gian tọa độ mới. Do Λ là ma trận đường chéo, người ta kết luận rằng sau phép chuyển đổi, mỗi chiều u_i mới có phương sai λ_i , và tất cả các hiệp phương sai là 0. Nói khác đi, trong không gian mới, phân phối chuẩn là theo trục ngang, không có thành phần xoay, nhưng nghiêng theo mỗi trục, tỉ lệ với giá trị riêng λ_i ; λ_i thể hiện phương sai theo chiều này.

Phương sai tổng cộng và phương sai tổng quát: định thức của ma trận hiệp phương sai là $\det(\Sigma) = \prod_{i=1}^d \lambda_i$. Vậy phương sai tổng quát của Σ là tích của các giá trị riêng của nó. Biết vết của ma trận vuông là bất biến với phép chuyển tương tự, như thay đổi cơ sở, người ta thấy phương sai tổng cộng $\text{var}(D)$ đối với bộ dữ liệu D là bất biến, tức $\text{var}(D) = \text{tr}(\Sigma) = \sum_{i=1}^d \sigma_i^2 = \sum_{i=1}^d \lambda_i = \text{tr}(\Lambda)$. Nói khác đi, $\sigma_1^2 + \dots + \sigma_d^2 = \lambda_1 + \dots + \lambda_d$.

Về mật độ chuẩn hai biến, xét thuộc tính độ dài là dài X_1 và độ rộng X_2 , như các biến ngẫu nhiên, người ta xác định biến ngẫu nhiên hai biến $X = (X_1 \ X_2)$. Giả sử X theo phân phối chuẩn hai biến, có thể ước lượng các tham số của nó từ mẫu. Trung bình mẫu là $\hat{\mu} = (5.843, \ 3.054)^T$. Ma trận hiệp phương sai mẫu $\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$.

Đồ thị của mật độ chuẩn hai biến như trong hình; nó cho thấy đường viền và các điểm dữ liệu.



Hình 4.7. Mật độ chuẩn hai biến và đường viền

Xét điểm $x_2 = (6.9, 3.1)^T$, có $x_2 - \hat{\mu} = \begin{pmatrix} 6.9 \\ 3.1 \end{pmatrix} - \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix} = \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}$.

Khoảng cách Mahalanobis giữa x_2 và $\hat{\mu}$ là: $(x_2 - \hat{\mu})^T \hat{\Sigma}^{-1} (x_2 - \hat{\mu})$

$$= (1.057 \ 0.046) \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}^{-1} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} = 1.701; \text{ trong khi khoảng}$$

$$\text{cách Oclit giữa chúng là } \|x_2 - \hat{\mu}\|^2 = (1.057 \ 0.046) \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} = 1.119.$$

Các giá trị riêng và các vector riêng tương ứng của $\hat{\Sigma}$ là: $\lambda_1 = 0.684$, $u_1 = (-0.997, 0.078)^T$, $\lambda_2 = 0.184$, $u_2 = (-0.078, -0.997)^T$. Hai vector riêng

này xác định các trục mới, mà ma trận hiệp phương sai là $\Lambda = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$.

Góc giữa trục ban đầu $e_1 = (1, 0)^T$ và u_1 xác định góc xoay đối với chuẩn đa biến: $\cos \theta = e_1^T u_1 = -0.997$; $\theta = \cos^{-1}(-0.997) = 175.5^\circ$. Hình vẽ mô tả các trục tọa độ mới và các phương sai mới. Theo trục ban đầu, các đường viền là xoay nhẹ theo góc 175.5° .

4.2. Các thuộc tính phạm trù

Đa số các chủ đề trên đề cập các phương pháp khai phá các thuộc tính số. Các thuộc tính phạm trù chỉ có giá trị kí hiệu, nên

nhiều phép toán số học không áp dụng. Tuy nhiên, với thuộc tính phạm trù, người ta tính tần suất của các giá trị.

4.2.1. Phân tích một biến

Giả sử dữ liệu gồm các giá trị của thuộc tính phạm trù đơn, X . Miền dữ liệu của X gồm m giá trị kí hiệu, $\text{dom}(X) = \{a_1, a_2, \dots, a_m\}$. Dữ liệu D là ma trận dữ liệu; kí hiệu,

$$D = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

trong đó mỗi $x_i \in \text{dom}(X)$.

4.2.1.1. Biến Bernoulli

Xét biến phạm trù X có miền giá trị $\{a_1, a_2\}$, tức $m = 2$. Người ta mô hình hóa X như biến ngẫu nhiên Bernoulli.

Định nghĩa: Biến ngẫu nhiên Bernoulli X nhận hai giá trị phân biệt, 1 và 0, theo ánh xạ $X(v) = 1$ nếu $v = a_i$; ngược lại, 0.

Hàm mật độ xác suất PMF của X là $P(X = x) = f(x) = p_1$ nếu $x = 1$; $f(x) = p_0$ nếu $x = 0$; trong đó p_0, p_1 là các tham số của phân phối, thỏa mãn dữ liệu $p_0 + p_1 = 1$. Như vậy trong p_0, p_1 chỉ có một là tham số tự do, nên người ta xem $p_1 = p$, và $p_0 = 1 - p$. Hàm PMF của biến ngẫu nhiên Bernoulli được viết lại là $P(X = x) = f(x) = p^x (1 - p)^{1-x}$.

Trung bình và phương sai:

Định nghĩa: Giá trị kì vọng của biến ngẫu nhiên Bernoulli X là $\mu = E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$.

Định nghĩa: Phương sai của X là $\sigma^2 = \text{var}(X) = E[X^2] - (E[X])^2 = (1^2 \cdot p + 0^2 \cdot (1 - p)) - p^2 = p(1 - p)$.

Trung bình mẫu và phương sai mẫu: Để ước lượng các tham số của biến Bernoulli X , giả sử mỗi điểm kí hiệu có ánh xạ với giá trị nhị phân. Vậy tập $\{x_1, x_2, \dots, x_n\}$ được giả thiết là mẫu ngẫu nhiên rút từ X , tức mỗi x_i là IID với X .

Trung bình mẫu là $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n_1}{n} = \hat{p}$; trong đó n_1 là số các điểm có $x_i = 1$ trong mẫu ngẫu nhiên, tức số các xuất hiện của kí hiệu a_1 . Lấy $n_0 = n - n_1$ thay cho số các điểm có $x_i = 0$ trong mẫu ngẫu nhiên, phương sai mẫu là $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{n_1}{n} (1 - \hat{p})^2 + \frac{n_0}{n} (0 - \hat{p})^2 = \hat{p}(1 - \hat{p})^2 + (1 - \hat{p})\hat{p}^2 = \hat{p}(1 - \hat{p})$. Phương sai mẫu cũng có thể rút ra từ định nghĩa $\sigma^2 = p(1 - p)$ khi thay p bằng $\hat{p}$.

Xét thuộc tính độ dài lá đài, X_1 . Giả sử hoa là *dài* nếu độ dài lá đài trong $[7, \infty)$, *ngắn* nếu trong $(-\infty, 7)$. Vậy X_1 có thể được xem như thuộc tính phạm trù với miền {dài, ngắn}. Từ mẫu quan sát với $n = 150$, người ta thấy có 13 hoa dài. Trung bình mẫu của X_1 là $\hat{\mu} = \hat{p} = 13/150 = 0.087$; và phương sai của nó là $\hat{\sigma}^2 = \hat{p}(1 - \hat{p}) = 0.087 \cdot (1 - 0.087) = 0.079$.

Phân phối nhị thức: Cho biến Bernoulli X , với các mẫu ngẫu nhiên kích thước n là $\{x_1, x_2, \dots, x_n\}$ lấy ra từ X . Giả sử N là biến ngẫu nhiên, thể hiện số lần xuất hiện của kí hiệu a_1 , tức giá trị $X = 1$, trong mẫu này.

Định nghĩa: Số lần xuất hiện kí hiệu a_1 , tức N , có phân phối nhị

$$\text{thức } f(N = n_1 | n, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1}.$$

Trong đó $\binom{n}{k}$ là tổ hợp chập k của n phần tử

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots 1} = \frac{n!}{k!(n-k)!}.$$

Sự thực N là tổng của n biến

độc lập x_i Bernoulli IID với X , tức $N = \sum_{i=1}^n x_i$. Do tính tuyến tính

của kì vọng, trung bình hay số kì vọng của việc xuất hiện kí hiệu a_1 là $\mu_N = E[N] = E[\sum_{i=1}^n x_i] = \sum_{i=1}^n E[x_i] = \sum_{i=1}^n p = np$. Do x_i độc lập, phương sai của N là $\sigma_N^2 = \text{var}(N) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n p(1-p) = np(1-p)$.

Có thể sử dụng tham số ước lượng $\hat{p} = 0.087$ để tính số kì vọng của số lần xuất hiện N của độ dài lá dài dài, thông quan phân phối nhị thức $E[N] = n\hat{p} = 150 \cdot 0.087 \approx 13$. Do p được ước lượng từ mẫu thông qua $\hat{p}$, người ta thấy số kì vọng về xuất hiện hoa dài gần với số xuất hiện thực sự. Tuy nhiên, điều quan trọng là người ta có thể tính phương sai về số lần xuất hiện: $\text{var}(N) = n\hat{p}(1-\hat{p}) = 150 \cdot 0.079 \approx 11.9$. Do kích thước mẫu tăng, phân phối nhị thức có xu hướng dần đến phân phối chuẩn với $\mu = 13$ và $\sigma = \sqrt{11.9} = 3.45$. Do vậy, với độ tin cậy lớn hơn 95%, có thể khẳng định rằng số xuất hiện của a_1 sẽ nằm trong khoảng $\mu \pm 2\sigma = [9.55, 16.45]$, với sự thật là đối với phân phối chuẩn, 95.45% mật độ xác suất nằm trong hai độ lệch chuẩn so với trung bình.

4.2.1.2. Biến Bernoulli đa biến

Xét trường hợp tổng quát, khi X là thuộc tính phạm trù, với miền $\{a_1, \dots, a_m\}$. Có thể mô hình hóa X như biến ngẫu nhiên Bernoulli m chiều, $X = (A_1, A_2, \dots, A_m)^T$, trong đó A_i là biến Bernoulli với tham số xác suất p_i của kí hiệu quan sát a_i . Do giả thiết X lấy một giá trị kí hiệu tại thời điểm, nếu $X = a_i$ thì $A_i = 1$ và các $A_j = 0$. Khoảng của biến X ngẫu nhiên là tập $\{0, 1\}^m$, với hạn chế rằng $X = a_i$ thì $X = e_i$; trong đó e_i là vector chuẩn cơ sở thứ i , $e_i \in \mathbb{R}^m$, với $e_i = (\underbrace{0, \dots, 0}_{i-1}, 1, \underbrace{0, \dots, 0}_{m-i})^T$.

Định nghĩa: Biến Bernoulli đa biến là tổng quát của biến Bernoulli hai biến, cho m biến, với $X(v) = e_i$ nếu $v = a_i$.

Khoảng của X bao gồm m giá trị vector phân biệt $\{e_1, e_2, \dots, e_m\}$, với PMF của X là $P(X = e_i) = f(e_i) = p_i$; trong đó p_i là xác suất của giá

trị quan sát a_i . Các tham số này cần thỏa mãn điều kiện $\sum_{i=1}^m p_i = 1$. Hàm PMF có thể viết lại $P(X = e_i) = f(e_i) = \prod_{j=1}^m p_j^{e_{ij}}$. Do $e_{ii} = 1$ và $e_{ij} = 0$, người ta có $f(e_i) = \prod_{j=1}^m p_j^{e_{ij}} = p_1^{e_{i1}} \dots p_i^{e_{ii}} \dots p_m^{e_{im}} = p_i$.

Xét thuộc tính độ dài lá đài X_1 với dữ liệu trong bảng. Người ta chia độ dài lá đài thành bốn khoảng độ dài bằng nhau, đặt tên cho khoảng. X_1 như thuộc tính phạm trù với miền $\{a_1 = \text{rất ngắn}, a_2 = \text{ngắn}, a_3 = \text{dài}, a_4 = \text{rất dài}\}$.

Bảng 4.2. Thuộc tính độ dài lá đài rời rạc

Khoảng	Miền	Số đếm
[4.3, 5.2]	Rất ngắn (a_1)	$N_1 = 45$
(5.2, 6.1]	Ngắn (a_2)	$N_2 = 50$
(6.1, 7.0]	Dài (a_3)	$N_3 = 43$
(7.0, 7.9]	Rất dài (a_4)	$N_4 = 12$

Mô hình hóa thuộc tính phạm trù X_1 như biến Bernoulli đa biến X , xác định như $X(v) = \begin{cases} e_1 = (1,0,0,0) & \text{ khi } v = a_1 \\ e_2 = (0,1,0,0) & \text{ khi } v = a_2 \\ e_3 = (0,0,1,0) & \text{ khi } v = a_3 \\ e_4 = (0,0,0,1) & \text{ khi } v = a_4 \end{cases}$.

Chẳng hạn điểm kí hiệu $x_1 = \text{ngắn} = a_2$ được thể hiện như vector $(0, 1, 0, 0)^T = e_2$.

Trung bình: Trung bình hay giá trị kì vọng của X là $\mu = E[X] =$

$$\sum_{i=1}^m e_i f(e_i) = \sum_{i=1}^m e_i p_i = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix} p_1 + \dots + \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \dots \\ p_m \end{pmatrix} = p.$$

Trung bình mẫu: Giả sử mỗi điểm kí hiệu $x_i \in D$ được ánh xạ sang biến $x_i = X(x_i)$. Bộ dữ liệu được ánh xạ $x_1, x_2, \dots, x_n$ giả thiết là mẫu ngẫu nhiên IID với X . Người ta có thể tính trung bình mẫu

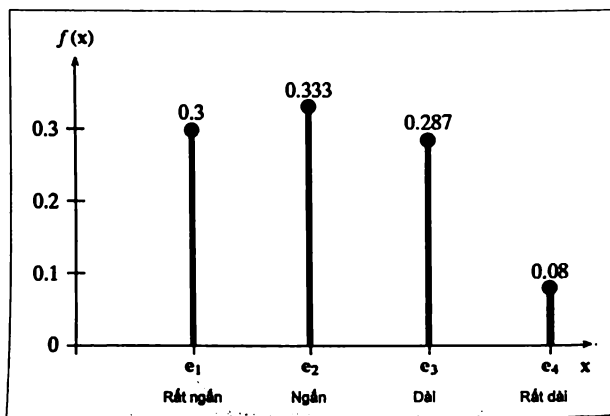
nhờ thay thế mật độ xác suất của $1/n$ tại mỗi điểm $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$,

$$= \sum_{i=1}^m \frac{n_i}{n} e_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \dots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \dots \\ \hat{p}_m \end{pmatrix} = \hat{p}; \text{ trong đó } n_i \text{ là số xuất hiện của}$$

giá trị vector e_i trong mẫu này, tương đương với số xuất hiện của kí hiệu a_i . Ngoài ra, $\sum_{i=1}^m n_i = n$, ứng với hiện tượng X chỉ có thể lấy m giá trị e_i phân biệt, và số đếm của mỗi giá trị cần thêm vào kích thước mẫu n .

Về trung bình mẫu, xét số đếm n_i quan sát đối với mỗi giá trị a_i , tức thuộc tính độ dài lá dài, như trong bảng dữ liệu. Do kích thước mẫu tổng cộng là $n = 150$, có thể ước lượng $\hat{p}_i$ như sau: $\hat{p}_1 = 45/150 = 0.3$; $\hat{p}_2 = 50/150 = 0.333$; $\hat{p}_3 = 43/150 = 0.287$; $\hat{p}_4 = 12/150 = 0.08$. Hàm PMF đối với X được vẽ như trong hình,

và trung bình mẫu là $\hat{\mu} = \hat{p} = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix}$.



Hình 4.8. Hàm mật độ xác suất của độ dài lá dài

Ma trận hiệp phương sai: Biến Bernoulli đa biến đơn giản là vectơ m biến Bernoulli. $X = (A_1, \dots, A_m)^T$, trong đó A_i là biến Bernoulli ứng với kí hiệu a_i . Thông tin hiệp phương sai - phương sai giữa các biến Bernoulli cho ma trận hiệp phương sai đối với X .

Trước tiên xét phương sai theo mỗi biến Bernoulli A_i . Theo công thức phương sai đơn biến, $\sigma_i^2 = \text{var}(A_i) = p_i(1 - p_i)$. Tiếp theo, xét hiệp phương sai giữa A_i và A_j . $\sigma_{ij} = E[A_i A_j] - E[A_i]E[A_j] = 0 - p_i p_j = -p_i p_j$, ứng với sự thật rằng $E[A_i A_j] = 0$ do A_i và A_j không đồng thời là 1, nên tích $A_i A_j = 0$. Điều quan trọng ở đây là độ kết hợp âm tỉ lệ với tích các giá trị trung bình đối với A_i và A_j .

Ma trận hiệp phương sai $m \times m$ là:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{1m} & \sigma_{2m} & \dots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \dots & -p_1 p_m \\ -p_1 p_2 & p_2(1-p_2) & \dots & -p_2 p_m \\ \dots & \dots & \dots & \dots \\ -p_1 p_m & -p_2 p_m & \dots & p_m(1-p_m) \end{pmatrix}.$$

Mỗi dòng có tổng là 0, tức $-p_1 p_1 - p_1 p_2 - \dots + p_1(1 - p_1) - \dots - p_1 p_m = p_1 - p_1 \sum_{j=1}^m p_j = p_1 - p_1 = 0$. Do Σ đối xứng, tổng cột cũng là 0. Xác

định P như ma trận đường chéo $m \times m$, $P = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & p_m \end{pmatrix}$, có thể viết

lại ma trận hiệp phương sai của X là $\Sigma = P - P^* P^T$.

Ma trận hiệp phương sai mẫu: Ma trận hiệp phương sai mẫu suy trực tiếp từ $\Sigma = P - P^* P^T$, được $\hat{\Sigma} = \hat{P} - \hat{P}^* \hat{P}^T$. Trong công thức, $\hat{p} = \bar{\mu} = (\bar{p}_1, \bar{p}_2, \dots, \bar{p}_m)^T$ là hàm mật độ xác suất thực nghiệm đối với X .

Quay lại thuộc tính độ dài lá dài rời rạc hóa, người ta có $\hat{p} = \bar{\mu} = (0.3, 0.333, 0.287, 0.08)^T$. Ma trận hiệp phương sai mẫu là

$$\hat{\Sigma} = \hat{P} - \hat{P} * \hat{P}^T$$

$$\begin{aligned}\hat{\Sigma} &= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} (0.3 \ 0.333 \ 0.287 \ 0.08) \\ &= \begin{pmatrix} 0.3 & 0 & 0 & 0 \\ 0 & 0.333 & 0 & 0 \\ 0 & 0 & 0.287 & 0 \\ 0 & 0 & 0 & 0.08 \end{pmatrix} - \begin{pmatrix} 0.09 & 0.1 & 0.086 & 0.024 \\ 0.1 & 0.111 & 0.096 & 0.027 \\ 0.086 & 0.096 & 0.082 & 0.023 \\ 0.024 & 0.027 & 0.023 & 0.006 \end{pmatrix} \\ &= \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix}\end{aligned}$$

Tổng theo dòng, hay cột, đều có giá trị 0.

Cần nhấn mạnh rằng khi mô hình hóa thuộc tính phạm trù X như biến Bernoulli đa biến, $X = (A_1, \dots, A_m)^T$, cho cấu trúc của ma trận trung bình và hiệp phương sai tường minh, người ta thu được kết quả tương tự như xét các giá trị ánh xạ $X(x_i)$ như ma trận dữ liệu nhị phân $n \times m$, và áp dụng các định nghĩa chuẩn về ma trận trung bình và hiệp phương sai từ phân tích thuộc tính số đa chiều. Đặc biệt, ánh xạ từ kí hiệu a_i sang vector nhị phân e_i là ý tưởng chính trong phân tích thuộc tính phạm trù.

Xét mẫu D với kích thước $n = 5$ về thuộc tính độ dài lá dài X_1 . Giả sử X_1 chỉ có hai giá trị phạm trù $\{\text{dài}, \text{ngắn}\}$. Người ta mô hình hóa X_1 như biến Bernoulli đa biến X_1 theo: $X_1(v) = e_1 = (1, 0)^T$ nếu $v = \text{dài}$, tức a_1 ; nếu $v = \text{ngắn}$, tức a_2 , thì $X_1(v) = e_2 = (0, 1)^T$. Trung bình mẫu là $\hat{p} = \hat{\mu} = (2/5, 3/5)^T = (0.4, 0.6)^T$; ma trận hiệp phương sai mẫu là

$$\hat{\Sigma} = \hat{P} - \hat{P} * \hat{P}^T = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} (0.4 \ 0.6) = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} - \begin{pmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{pmatrix}$$

$$= \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}. \text{ Để cho thấy người ta sẽ thu được cùng kết quả}$$

nờ phân tích số chuẩn, ánh xạ thuộc tính phạm trù X sang hai thuộc tính Bernoulli A_1 và A_2 , ứng với các kí hiệu *dài* và *ngắn*. Bộ dữ liệu ánh xạ như trong bảng.

Bảng 4.3. Bộ dữ liệu phạm trù; bộ dữ liệu nhị phân được ánh xạ; và bộ dữ liệu quy tâm

	X
x_1	Ngắn
x_2	Ngắn
x_3	Dài
x_4	Ngắn
x_5	Dài

	A_1	A_2
x_1	0	1
x_2	0	1
x_3	1	0
x_4	0	1
x_5	1	0

	Z_1	Z_2
z_1	-0.4	0.4
z_2	-0.4	0.4
z_3	0.6	-0.6
z_4	-0.4	0.4
z_5	0.6	-0.6

Trung bình mẫu đơn giản là $\bar{\mu} = \frac{1}{5} \sum_{i=1}^5 x_i = 0.2^* (2, 3)^T = (0.4, 0.6)^T$.

Tiếp theo, lấy tâm bộ dữ liệu bằng cách trừ giá trị thuộc tính một lượng bằng giá trị trung bình. Trong bảng cho thấy thuộc tính Z_1 là thuộc tính A_1 quy tâm. Có thể tính ma trận hiệp phương sai nhờ

tích trong, trên các vectơ cột quy tâm: $\sigma_1^2 = \frac{1}{5} Z_1^T Z_1 = 1.2/5 = 0.24$;

$\sigma_2^2 = \frac{1}{5} Z_2^T Z_2 = 1.2/5 = 0.24$; $\sigma_{12} = \frac{1}{5} Z_1^T Z_2 = -1.2/5 = -0.24$. Do vậy,

ma trận hiệp phương sai mẫu là $\hat{\Sigma} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$, khớp với kết

quả sử dụng tiếp cận mô hình hóa Bernoulli đa biến.

Phân phối đa thức: Phân phối đa thức¹ là cho số lần xuất hiện. Cho biến Bernoulli đa biến X và mẫu ngẫu nhiên $\{x_1, x_2, \dots, x_n\}$ lấy từ X . Gọi N_i là biến ngẫu nhiên tương ứng với số xuất hiện kí hiệu a_i trong mẫu này, và lấy $N = (N_1, N_2, \dots, N_m)^T$ là biến vectơ ngẫu nhiên ứng với phân phối kết hợp của số xuất hiện của toàn thể các kí hiệu. Vậy N có phân phối đa thức, $f(N = (n_1, n_2, \dots, n_m) | p)$

$$= \frac{n!}{n_1! * n_2! * \dots * n_m!} \prod_{i=1}^m p_i^{n_i}$$
. Trong công thức này, $\frac{n!}{n_1! * n_2! * \dots * n_m!}$ là

¹ Multinomial distribution: phân phối đa thức.

số cách chọn n_i xuất hiện kí hiệu a_i từ mẫu kích thước n , và $\sum_{i=1}^m n_i = n$.

Ma trận trung bình và hiệp phương sai của N là gấp n lần ma trận trung bình và hiệp phương sai của X . Tức $\mu_N = E[N] = nE[X] = n^* \mu = n^* p = (np_1 \dots np_m)^T$ và $\Sigma_N = n^* (p - p^* p^T) =$

$$\begin{pmatrix} np_1(1-p_1) - np_1p_2 \dots - np_1p_m \\ -np_1p_2 \quad np_2(1-p_2) \dots - np_2p_m \\ \dots \\ -np_1p_m \quad -np_2p_m \dots np_m(1-p_m) \end{pmatrix}.$$

Tương tự, ma trận trung bình mẫu và ma trận hiệp phương sai mẫu đối với N là $\hat{\mu}_N = n\hat{p}$ và $\Sigma_N = n(\hat{p} - \hat{p}^* \hat{p}^T)$.

4.2.2. Phân tích hai biến

4.2.2.1. Giới thiệu về phân tích hai biến

Giả sử dữ liệu gồm hai thuộc tính phạm trù, X_1 và X_2 , với $\text{dom}(X_1) = \{a_{11}, a_{12}, \dots, a_{1m_1}\}$ và $\text{dom}(X_2) = \{a_{21}, a_{22}, \dots, a_{2m_2}\}$. Người ta có n điểm phạm trù có dạng $x_i = (x_{i1}, x_{i2})^T$ mà $x_{ij} \in \text{dom}(X_j)$. Bộ dữ liệu là ma trận dữ liệu kí hiệu $n \times 2$.

$$D = \begin{pmatrix} X_1 & X_2 \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Có thể mô hình hóa X_1 và X_2 như các biến Bernoulli đa biến X_1, X_2 với m_1, m_2 chiều. Hàm mật độ xác suất đối với X_1 và X_2 là $P(X_1 = e_{1i}) = f_1(e_{1i}) = p_i^1 = \prod_{k=1}^{m_1} (p_i^1)^{e_{1ik}}$ và $P(X_2 = e_{2j}) = f_1(e_{2j}) = p_j^2 = \prod_{k=1}^{m_2} (p_j^2)^{e_{2jk}}$; trong đó e_{1i} là vectơ cơ sở chuẩn thứ i trong R^{m_1} , đối với thuộc tính X_1 , mà thành phần thứ k của nó là e_{1ik} ; tham số p_i^1 là xác suất quan sát kí hiệu a_{1i} . Điều kiện về xác suất $\sum_{i=1}^{m_1} p_i^1 = 1 = \sum_{j=1}^{m_2} p_j^2$.

Phân phối kết hợp của X_1 và X_2 được mô hình hóa theo biến vectơ $d' = m_1 + m_2$ chiều, $X = (X_1 \ X_2)^T$, xác định theo $X((v_1, v_2)^T) = \begin{pmatrix} X_1(v_1) \\ X_2(v_2) \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2j} \end{pmatrix}$, đảm bảo $v_1 = a_{1i}$ và $v_2 = a_{2j}$. Khoảng của X gồm $m_1 \cdot m_2$ cặp khác nhau của các giá trị vectơ $\{ (e_{1i}, e_{2j})^T \}$, $i = 1..m_1$; $j = 1..m_2$. Hàm PMF của X là $P(X = (e_{1i}, e_{2j})^T) = f(e_{1i}, e_{2j}) = p_{ij} = \prod_{r=1}^{m_1} \prod_{s=1}^{m_2} p_{ij}^{e_{1r} \cdot e_{2s}}$; trong đó p_{ij} là xác suất quan sát cặp kí hiệu (a_{1i}, a_{2j}) . Các tham số xác suất này cần thỏa mãn các điều kiện $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$. Hàm PMF đối với x được biểu diễn như ma trận

$$m_1 \cdot m_2: P_{12} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1m_2} \\ p_{21} & p_{22} & \dots & p_{2m_2} \\ \dots & \dots & \dots & \dots \\ p_{m_1 1} & p_{m_1 2} & \dots & p_{m_1 m_2} \end{pmatrix}.$$

Xét thuộc tính độ dài lá đài, X_1 . Người ta cũng rời rạc hóa thuộc tính độ rộng lá đài, X_2 . $\text{Dom}(X_1) = \{a_{11} = \text{rất ngắn}, a_{12} = \text{ngắn}, a_{13} = \text{dài}, a_{14} = \text{rất dài}\}$; $\text{dom}(X_2) = \{a_{21} = \text{ngắn}, a_{22} = \text{trung bình}, a_{23} = \text{dài}\}$. Điểm kí hiệu $x = (\text{ngắn}, \text{dài}) = (a_{12}, a_{23})$ được ánh xạ sang vectơ $X(x) = (e_{12} \ e_{23})^T = (0, 1, 0, 0 \mid 0, 0, 1)^T \in R^7$.

Bảng 4.4. Thuộc tính độ rộng lá đài rời rạc

Khoảng	Miền	Số đếm
[2.0, 2.8]	Ngắn, tức a_1	47
(2.8, 3.6]	Trung bình, tức a_2	88
(3.6, 4.4]	Dài, tức a_3	15

Trong đẳng thức trên, e_{12} là vectơ cơ sở chuẩn thứ hai trong R_4 đối với x_1 , và e_{23} là vectơ cơ sở chuẩn thứ ba trong R_3 đối với X_2 .

Trung bình: Trung bình hai biến là tổng quát hóa của trường hợp một biến, $\mu = E[X] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$; trong

đó $\mu_i = p_i = (p_{i1}, \dots, p_{im_i})^T$, là vector trung bình của X_i . Các vector p_i là hàm mật độ xác suất đối với X_i .

Trung bình mẫu: Trung bình mẫu được tính bằng cách thay mật độ xác suất $1/n$ tại mỗi điểm.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \left(\frac{\sum_{i=1}^{m_1} n_{i1} e_{1i}}{\sum_{i=1}^{m_2} n_{i2} e_{2i}} \right) = \frac{1}{n} \begin{pmatrix} n_{11} \\ \dots \\ n_{1m_1} \\ n_{21} \\ \dots \\ n_{2m_2} \end{pmatrix} = \begin{pmatrix} \hat{p}_{11} \\ \dots \\ \hat{p}_{1m_1} \\ \hat{p}_{21} \\ \dots \\ \hat{p}_{2m_2} \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix},$$

trong đó n_{ij} là tần suất quan sát của kí hiệu a_{ij} trong mẫu kích thước n , và $\bar{\mu}_i = \bar{p}_i = (p_{i1}, \dots, p_{im_i})^T$ là vector trung bình mẫu đối với X_i , cũng là PMF thực nghiệm đối với thuộc tính X_i .

Ma trận hiệp phương sai: Ma trận hiệp phương sai đối với X là ma trận $d'^*d' = (m_1 + m_2) \times (m_1 + m_2)$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$; trong đó Σ_{ii} là ma trận hiệp phương sai $m_i \times m_i$ đối với X_i . $\Sigma_{ii} = P_i - p_i p_i^T$ và P_i là ma trận đường chéo p_i . Ma trận Σ_{12} là ma trận hiệp phương sai $m_1 \times m_2$ giữa biến X_1 và X_2 , $\Sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)^T] = E[X_1 X_2^T] - E[X_1]E[X_2]^T = P_{12} - \mu_1 \mu_2^T = P_{12} - p_1 p_2^T = \begin{pmatrix} p_{11} - p_{11}^2 & \dots & p_{1m_2} - p_{11} p_{2m_2} \\ \dots & & \\ p_{m_1 1} - p_{1m_1} p_{21} & \dots & p_{m_1 m_2} - p_{1m_1} p_{2m_2} \end{pmatrix}$. P_{12} thể hiện PMF kết hợp đôi với X .

Mỗi dòng và mỗi cột của Σ_{12} có tổng là 0, ứng với việc tổng các hàm mật độ kết hợp trên tất cả các giá trị của X_2 có phân phối lề của X_1 , và tổng các hàm mật độ kết hợp trên tất cả các giá trị của X_1 có phân phối lề của X_2 .

Ma trận hiệp phương sai mẫu: ma trận hiệp phương sai là $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{pmatrix}$; trong đó $\hat{\Sigma}_{11} = \hat{P}_1 - \hat{p}_1 \hat{p}_1^T$, $\hat{\Sigma}_{12} = \hat{P}_{12} - \hat{p}_1 \hat{p}_2^T$, $\hat{P}_i$ là ma trận đường chéo $\hat{p}_i$; $\hat{p}_i$ xác định hàm mật độ xác suất thực nghiệm đối với X_i . $\hat{P}_{12}$ là PMF kết hợp thực nghiệm đối với X_1 và X_2 , tính theo $\hat{P}_{12}(i, j) = \hat{f}(e_{1i}, e_{2j}) = \frac{1}{n} \sum_{k=1}^n I_{ij}(x_k) = \frac{n_{ij}}{n} = \hat{p}_{ij}$; I_{ij} là biến. Xác định $I_{ij}(x_k) = 1$ nếu $x_{k1} = e_{1i}$ và $x_{k2} = e_{2j}$; ngược lại là 0. Lấy tổng của $I_{ij}(x_k)$ trên tất cả n điểm trong mẫu, thu được số các xuất hiện n_{ij} của cặp kí hiệu (a_{1i}, a_{2j}) trong mẫu này. Với ma trận hiệp phương sai thuộc tính chéo $\hat{\Sigma}_{12}$, cần ước lượng số các tham số bậc hai. Tức người ta cần có số đếm n_{ij} tin cậy để ước lượng các tham số p_{ij} , đối với tổng cộng $O(m_1 * m_2)$ tham số cần được ước lượng; đó là khó khăn nếu các tham số phạm trù có nhiều kí hiệu. Mặt khác, ước lượng $\hat{\Sigma}_{11}$ đòi hỏi ước lượng tham số m_1 và m_2 , ứng với p^1 và p^2 . Nói chung, việc tính đòi hỏi ước lượng $m_1 m_2 + m_1 + m_2$ tham số.

Xét các thuộc tính phạm trù hai biến X_1, X_2 . Theo số đếm xuất hiện đối với mỗi giá trị về độ rộng dải hoa, người ta có

$$\hat{\mu}_1 = \hat{p}_1 = \begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix}, \hat{\mu}_2 = \hat{p}_2 = \frac{1}{150} \begin{pmatrix} 47 \\ 88 \\ 15 \end{pmatrix} = \begin{pmatrix} 0.313 \\ 0.587 \\ 0.1 \end{pmatrix}. \text{ Do vậy trung bình}$$

đối với $X = (X_1, X_2)^T$ là $\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = (0.3, 0.333, 0.287, 0.08 \mid 0.313, 0.587, 0.1)^T$.

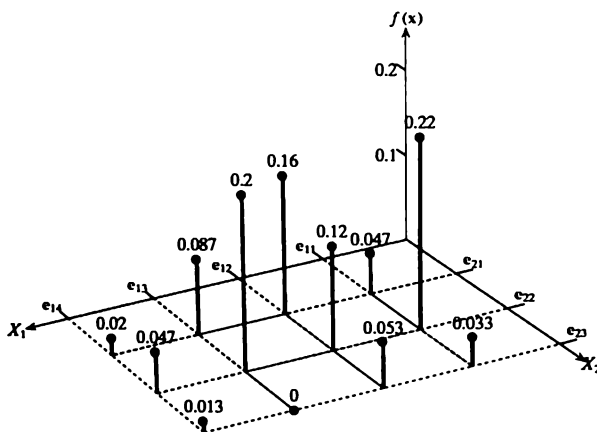
$$\hat{\Sigma}_{11} = \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix},$$

$$\hat{\Sigma}_{22} = \begin{pmatrix} 0.215 & -0.184 & -0.031 \\ -0.184 & 0.242 & -0.059 \\ -0.031 & -0.059 & 0.09 \end{pmatrix}$$

Tiếp theo, sử dụng các số đếm quan sát để có PMF kết hợp thực nghiệm đối với X_1 , X_2 , dùng bảng sau.

Bảng 4.5. Số đếm quan sát n_{ij} về độ dài lá đài và độ rộng lá đài

X_2		Ngắn (e_{21})	Vừa (e_{22})	Dài (e_{23})
X_1	Rất ngắn (e_{11})	7	33	5
	Ngắn (e_{12})	24	18	8
	Dài (e_{13})	13	30	0
	Rất dài (e_{14})	3	7	2



Hình 4.9. Hàm mật độ xác suất kết hợp thực nghiệm, đối với độ rộng và độ dài lá đài

Nhờ các xác suất này, thu được $E[X_1 X_2 T] =$

$$\hat{P}_{12} = \frac{1}{150} \begin{pmatrix} 7 & 33 & 5 \\ 24 & 18 & 8 \\ 13 & 30 & 0 \\ 3 & 7 & 2 \end{pmatrix} = \begin{pmatrix} 0.047 & 0.22 & 0.033 \\ 0.16 & 0.12 & 0.053 \\ 0.087 & 0.2 & 0 \\ 0.02 & 0.047 & 0.013 \end{pmatrix}. \text{ Ngoài ra, có ma trận}$$

hiệp phương sai mẫu: $E[X_1]E[X_2] = \hat{\mu}_1\hat{\mu}_2^T = \hat{p}_1\hat{p}_2^T =$

$$\begin{pmatrix} 0.3 \\ 0.333 \\ 0.287 \\ 0.08 \end{pmatrix} (0.313 \ 0.587 \ 0.1) = \begin{pmatrix} 0.094 & 0.176 & 0.03 \\ 0.104 & 0.196 & 0.033 \\ 0.09 & 0.168 & 0.029 \\ 0.025 & 0.047 & 0.008 \end{pmatrix}.$$

Có thể tính ma trận hiệp phương sai mẫu giữa các thuộc tính $\hat{\Sigma}_{12}$ đối với X_1 và X_2 theo $\hat{\Sigma}_{12} = \hat{P}_{12} - \hat{p}_1\hat{p}_2^T =$

$$\begin{pmatrix} -0.047 & 0.044 & 0.003 \\ 0.056 & -0.076 & 0.02 \\ -0.003 & 0.032 & -0.029 \\ -0.005 & 0 & 0.005 \end{pmatrix}.$$

Người ta có thể quan sát thấy mỗi dòng và cột trong $\hat{\Sigma}_{12}$ có tổng là 0. Đặt tất cả với nhau, người ta thu được ma trận hiệp phương sai mẫu $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{pmatrix} =$

$$\begin{pmatrix} \begin{pmatrix} 0.21 & -0.1 & -0.086 & -0.024 \\ -0.1 & 0.222 & -0.096 & -0.027 \\ -0.086 & -0.096 & 0.204 & -0.023 \\ -0.024 & -0.027 & -0.023 & 0.074 \end{pmatrix} & \begin{pmatrix} -0.047 & 0.044 & 0.003 \\ 0.056 & -0.076 & 0.02 \\ -0.003 & 0.032 & -0.029 \\ -0.005 & 0 & 0.005 \end{pmatrix} \\ \begin{pmatrix} -0.47 & 0.056 & -0.003 & -0.005 \\ 0.044 & -0.076 & 0.032 & 0 \\ 0.003 & 0.02 & -0.029 & 0.005 \end{pmatrix} & \begin{pmatrix} 0.215 & -0.184 & -0.031 \\ -0.184 & 0.242 & -0.059 \\ -0.031 & -0.059 & 0.09 \end{pmatrix} \end{pmatrix}.$$

4.2.2.2. Phụ thuộc về thuộc tính

Việc thử nghiệm về độc lập của hai biến phạm trù ngẫu nhiên X_1, X_2 được thực hiện qua *phân tích bảng ngẫu sinh*¹. Ý tưởng chính là thiết lập khung thử nghiệm giả thuyết, mà giả thuyết rỗng H_0 là X_1 và X_2 độc lập, và giả thuyết khác là X_1 và X_2 phụ thuộc. Người ta tính giá trị của thống kê χ^2 dưới giả thuyết rỗng. Tùy vào giá trị p , hoặc người ta chấp nhận hoặc loại bỏ giả thuyết rỗng.

¹ Contingency table analysis: phân tích bảng ngẫu sinh.

Bảng ngẫu sinh:

Định nghĩa: Bảng ngẫu sinh đối với X_1 và X_2 là ma trận $m_1 \times m_2$ của các số đếm quan sát n_{ij} đối với tất cả các cặp giá trị (e_{1i}, e_{2j}) trong mẫu đã biết với kích thước n , xác định theo $N_{12} = n$. $\hat{P}_{12}$.

$$\text{Bảng ngẫu sinh được chi tiết } N_{12} = n. \quad \hat{P}_{12} = \begin{pmatrix} n_{11} & n_{12} & n_{1m_2} \\ n_{21} & n_{22} & n_{2m_2} \\ n_{m_1 1} & n_{m_1 2} & n_{m_1 m_2} \end{pmatrix},$$

trong đó $\hat{P}_{12}$ là PMF kết hợp thực nghiệm đối với X_1 và X_2 . Bảng ngẫu sinh được tăng lên vào số đếm lề dòng và cột, theo $N_1 = n^* \hat{p}_1 = (n^1_1, \dots, n^1_{m_1})^T$, $N_2 = n^* \hat{p}_2 = (n^2_1, \dots, n^2_{m_2})^T$. Các mục dòng và cột lề và kích thước mẫu thỏa mãn ràng buộc sau: $n^1_i = \sum_{j=1}^{m_2} n_{ij}$; $n^2_j = \sum_{i=1}^{m_1} n_{ij}$; $n = \sum_{i=1}^{m_1} n^1_i = \sum_{j=1}^{m_2} n^2_j = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$. Cả N_1 và N_2 đều có phân phối nhị thức với tham số $p_1 = (p^1_1, \dots, p^1_{m_1})$ và $p_2 = (p^2_1, \dots, p^2_{m_2})$. Ngoài ra, N_{12} cũng có phân phối nhị thức với tham số $P_{12} = \{p_{ij}\}$.

Về bảng ngẫu sinh, xét bảng trước đây đã cho thấy các số đếm quan sát đối với độ dài lá đài và độ rộng lá đài rời rạc hóa, X_1, X_2 . Việc tăng số đếm lề hàng và cột và kích thước mẫu cho bảng kết quả sau.

Bảng 4.6. Bảng ngẫu sinh về độ dài lá đài và độ rộng lá đài.

X_2		Ngắn (a_{21})	Vừa (a_{22})	Dài (a_{23})	Đếm dòng
X_1	Rất ngắn (a_{11})	7	33	5	$n^1_1 = 45$
	Ngắn (a_{12})	24	18	8	$n^1_2 = 50$
	Dài (a_{13})	13	30	0	$n^1_3 = 43$
	Rất dài (a_{14})	3	7	2	$n^1_4 = 15$
	Đếm cột	$n^2_1 = 47$	$n^2_2 = 88$	$n^2_3 = 15$	$n = 150$

Kiểm định giả thuyết và thống kê χ^2 theo giả thuyết rỗng về X_1 và X_2 độc lập, hàm mật độ xác suất kết hợp của chúng là $\bar{p}_{ij} = \bar{p}_i^1 * \bar{p}_j^2$. Theo giả thiết độc lập, tần suất kì vọng đối với mỗi

cặp các giá trị là $e_{ij} = n * \bar{p}_{ij} = n * \bar{p}_i^1 * \bar{p}_j^2 = n \frac{n_i^1}{n} * \frac{n_j^2}{n} = \frac{n_i^1 * n_j^2}{n}$.

Tuy nhiên, từ mẫu người ta đã thấy được tần suất quan sát về mỗi cặp giá trị n_{ij} . Cần xác định sai khác có ý nghĩa giữa các tần suất quan sát với thực nghiệm: (i) nếu có sai khác có ý nghĩa, người ta chấp nhận giả thuyết rỗng, tức các thuộc tính là độc lập; (ii) nếu sai khác không có ý nghĩa, cần loại bỏ giả thuyết rỗng, tức người ta kết luận các thuộc tính là phụ thuộc.

Thống kê χ^2 lượng hóa sai khác giữa số đếm quan sát với số đếm thực nghiệm đối với mỗi cặp các giá trị. Nó được xác định

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \text{ Như vậy, cần xác định xác suất về việc}$$

đạt được giá trị χ^2 đã tính. Nói chung việc này khó khi không hiểu phân phối mẫu của thống kê đã biết. Đối với thống kê χ^2 , người ta đã biết phân phối mẫu theo hàm mật độ χ^2 , với bậc q tự do: $f(x | q)$

$$= \frac{1}{2^{\frac{q}{2}} \Gamma(\frac{q}{2})} x^{\frac{q}{2}-1} e^{-\frac{x}{2}}, \text{ trong đó hàm } \Gamma \text{ được xác định } \Gamma(k > 0) =$$

$\int_0^\infty x^{k-1} e^{-x} dx$. Bậc tự do q thể hiện số các tham số độc lập. Trong bảng ngẫu sinh có $m_1 * m_2$ số đếm quan sát n_{ij} . Tuy nhiên mỗi dòng i và mỗi cột j cần có tổng là n_i^1, n_j^2 . Ngoài ra, tổng của lề hàng và cột cần thêm n ; vậy người ta cần loại bỏ $m_1 + m_2$ tham số khỏi số các tham số độc lập. Tuy nhiên, do loại bỏ $n_{m_1 m_2}$ hai lần, nên tổng các bậc tự do là $q = |\text{dom}(X_1)| * |\text{dom}(X_2)| - (|\text{dom}(X_1)| + |\text{dom}(X_2)|) + 1 = m_1 * m_2 - m_1 - m_2 + 1 = (m_1 - 1)(m_2 - 1)$.

Giá trị p :

Định nghĩa: Giá trị p của thống kê θ là xác suất đạt được giá trị ít nhất ở mức cực đại như giá trị quan sát z , theo giả thuyết rỗng, được xác định giá trị $p(z) = P(\theta \geq z) = 1 - F(\theta)$.

Hàm $F(\theta)$ là phân phối xác suất tích lũy đối với thống kê. Giá trị p cho độ đo về các giá trị quan sát có ý nghĩa¹ của thống kê. Nếu giá trị quan sát nằm trong miền xác suất thấp, thì giá trị này là càng có ý nghĩa. Nói chung, giá trị p càng thấp, giá trị quan sát càng có ý nghĩa, và càng có cơ sở để loại trừ giả thuyết rỗng. Giả thuyết rỗng bị loại bỏ, nếu giá trị p thấp hơn ngưỡng ý nghĩa, α . Chẳng hạn nếu $\alpha = 0.01$, người ta có thể loại bỏ giả thuyết rỗng, khi giá trị $p \leq \alpha$. Đối với α đã biết, giá trị thống kê của phép thử z , với giá trị p , mà giá trị $p(z) = \alpha$, được gọi là giá trị nguy hiểm². Phép thử khác để loại bỏ giả thuyết rỗng là kiểm tra $\chi^2 > z$, do trong trường hợp này, giá trị p của giá trị χ^2 quan sát được bị α chặn, tức giá trị $p(\chi^2) \leq$ giá trị $p(z) = \alpha$. Giá trị $1 - \alpha$ được gọi là mức tin cậy³.

Xét bảng ngẫu sinh về độ dài và độ rộng lá dài, đã nêu. Người ta tính số đếm kì vọng; các số đếm hiện trong bảng sau. Chẳng hạn $e_{11} = n_{11} \cdot n_{21} / n = 45 \cdot 47 / 150 = 14.1$.

Bảng 4.7. Số đếm kì vọng

X_2		Ngắn (a_{21})	Vừa (a_{22})	Dài (a_{23})
X_1	Rất ngắn (a_{11})	14.1	26.4	4.5
	Ngắn (a_{12})	15.67	29.33	5.0
	Dài (a_{13})	13.47	25.23	4.3
	Rất dài (a_{14})	3.76	7.04	1.2

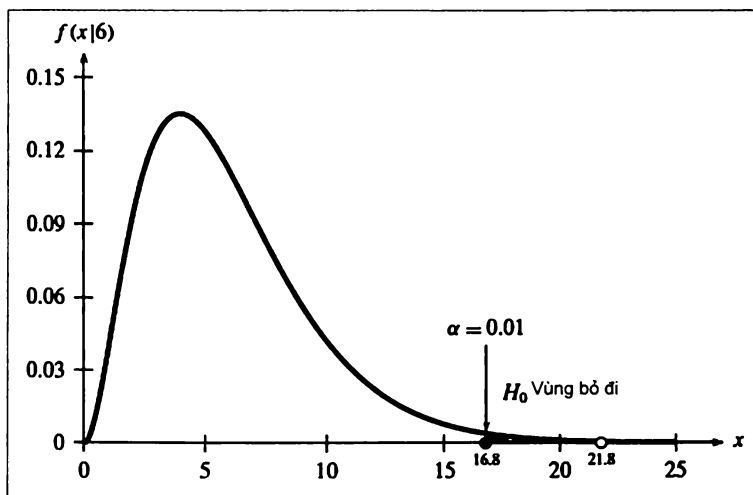
Tính thống kê χ^2 , được $\chi^2 = 21.8$. Ngoài ra, số bậc tự do là $q = (m_1 - 1)(m_2 - 1) = 3 \cdot 2 = 6$. Đồ thị hàm mật độ χ^2 với 6 bậc tự do thể hiện trong hình vẽ. Từ phân phối χ^2 tích lũy, người ta thu được giá trị $p(21.8) = 1 - F(21.8 | 6) = 1 - 0.9987 = 0.0013$. Với mức ý nghĩa $\alpha = 0.01$, người ta loại bỏ giả thuyết rỗng, do giá trị lớn của thống kê χ^2 là có ý nghĩa. Ngoài ra, tại mức ý nghĩa 0.01, giá trị nguy hiểm của thống kê là $z = F^{-1}(1 - 0.01 | 6) = 16.81$. Giá trị nguy

¹ Surprising value: giá trị đáng ngạc nhiên, tức có ý nghĩa.

² Critical value: giá trị nguy hiểm.

³ Confidence level: mức tin cậy.

hiếm này cũng thấy trên đồ thị hình vẽ, cho thấy rõ rằng giá trị quan sát 21.8 thuộc vào miền loại bỏ, do $21.8 > z = 16.81$. vậy người ta loại bỏ giả thuyết rằng độ rộng và độ dài lá đài là độc lập, và chấp nhận giả thuyết khác, rằng chúng phụ thuộc.



Hình 4.10. Phân phối χ^2 , $q = 6$

4.2.3. Phân tích đa biến

4.2.3.1. Giới thiệu về phân tích đa biến

Giả sử bộ dữ liệu có d thuộc tính phạm trù X_i , $\text{dom}(X_i) = \{a_{i1}, \dots, a_{im_i}\}$. Có n điểm phạm trù có dạng $x_i = (x_{i1}, \dots, x_{id})^T$, mà $x_{ij} \in \text{dom}(X_j)$. Bộ dữ liệu là ma trận kí hiệu $n \times d$:

$$D = \begin{pmatrix} X_1 & X_2 & \dots & X_d \\ x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & & \\ \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}$$

Mỗi thuộc tính X_i được mô hình hóa như biến Bernoulli X_i đa biến m_i chiều, và phân phối kết hợp của chúng được mô hình hóa

như biến vector d' chiều, $d' = \sum_{j=1}^d m_j$, $X = (X_1, \dots, X_d)^T$. Mỗi điểm dữ liệu phạm trù $v = (v_1, \dots, v_d)^T$ được biểu diễn như vector nhị phân d'

chiều. $X(v) = \begin{pmatrix} X_1(v_1) \\ \dots \\ X_d(v_d) \end{pmatrix} = \begin{pmatrix} e_{1k_1} \\ \dots \\ e_{dk_d} \end{pmatrix}$, mà $v_i = a_{ik_i}$ là kí hiệu thứ k_i của X_i .

Ở đây e_{ik_i} là vector cơ sở chuẩn thứ k_i trong R^{m_i} .

Trung bình: Tổng quát hóa trường hợp hai biến, đối với nhiều biến, trung bình và trung bình mẫu đối với X là:

$$\mu = E[X] = \begin{pmatrix} \mu_1 \\ \dots \\ \mu_d \end{pmatrix} = \begin{pmatrix} p_1 \\ \dots \\ p_d \end{pmatrix} \quad \hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \dots \\ \hat{\mu}_d \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \dots \\ \hat{p}_d \end{pmatrix};$$

trong đó $p_i = (p'_1, \dots, p'_m)^T$ là PMF đối với X_i , và $\hat{p}_i = (\hat{p}'_1, \dots, \hat{p}'_m)^T$ là PMF thực nghiệm đối với X_i .

Ma trận hiệp phương sai: Ma trận hiệp phương sai đối với X và các ước lượng của nó từ mẫu là ma trận $d' \times d'$:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1d} \\ \Sigma_{12}^T & \Sigma_{22} & & \Sigma_{2d} \\ \dots & \dots & \dots & \dots \\ \Sigma_{1d}^T & \Sigma_{2d}^T & & \Sigma_{dd} \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} & \dots & \hat{\Sigma}_{1d} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} & & \hat{\Sigma}_{2d} \\ \dots & \dots & \dots & \dots \\ \hat{\Sigma}_{1d}^T & \hat{\Sigma}_{2d}^T & & \hat{\Sigma}_{dd} \end{pmatrix}; \text{ trong đó } d' = \sum_{j=1}^d m_j,$$

và Σ_y là ma trận hiệp phương sai $m_i \times m_j$ đối với thuộc tính X_i và X_j ;

$\hat{\Sigma}_y$ là ước lượng của nó. $\Sigma_{ij} = P_{ij} - p_i p_j^T$ và $\hat{\Sigma}_{ij} = \hat{P}_{ij} - \hat{p}_i \hat{p}_j^T$; ở đây P_{ij} là PMF kết hợp và $\hat{P}_{ij}$ là PMF kết hợp thực nghiệm đối với X_i, X_j .

Về phân tích đa biến, xét tập con 3 chiều của bộ dữ liệu về hoa diên vĩ, với thuộc tính đã rời rạc hóa X_1 ứng với *độ dài lá dài*, X_2 *độ rộng lá dài* và thuộc tính phạm trù X_3 về *lớp*. Miền của X_1 và X_2 đã thể hiện trong các thí dụ trước. Có $\text{dom}(X_3) = \{\text{nhiều màu, mượt, có hoa}^1\}$. Mỗi giá trị X_3 xuất hiện 50 lần.

¹ Versicolor, setosa, virginica: tên riêng các hoa diên vĩ nhiều màu, mượt, có hoa.

Điểm x phạm trù $x = (\text{ngắn, vừa, nhiều màu})$ được mô hình hóa như vector $X(x) = \begin{pmatrix} e_{12} \\ e_{22} \\ e_{31} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \\ 1 & 0 & 0 & \end{pmatrix} \in R^{10}$. Theo kết quả thí dụ

trước và biết mỗi giá trị trong $\text{dom}(X_3)$ xuất hiện 50 lần trong mẫu $n = 150$, trung bình mẫu là $\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} 0.3 & 0.333 & 0.287 & 0.08 \\ 0.313 & 0.587 & 0.1 & \\ 0.33 & 0.33 & 0.33 & \end{pmatrix}$.

Sử dụng $\hat{p}_3 = (0.33, 0.33, 0.33)^T$, có thể tính ma trận hiệp phương

sai mẫu đối với X_3 : $\hat{\Sigma}_{33} = \begin{pmatrix} 0.222 & -0.111 & -0.111 \\ -0.111 & 0.222 & -0.111 \\ -0.111 & -0.111 & 0.222 \end{pmatrix}$. Sử dụng công

thức tính $\hat{\Sigma}_y$, thu được $\hat{\Sigma}_{13} = \begin{pmatrix} -0.067 & 0.16 & -0.093 \\ 0.082 & -0.038 & -0.044 \\ 0.011 & -0.096 & 0.084 \\ -0.027 & -0.027 & 0.053 \end{pmatrix}$ và

$\hat{\Sigma}_{23} = \begin{pmatrix} 0.076 & -0.098 & 0.022 \\ -0.042 & 0.044 & -0.002 \\ -0.033 & 0.053 & -0.02 \end{pmatrix}$. Kết hợp các ước lượng trung bình

mẫu $\hat{\Sigma}_y$ được ma trận hiệp phương sai mẫu 10×10 .

4.2.3.2. Phân tích ngẫu sinh nhiều đường

Để phân tích phụ thuộc *nhiều đường*¹, trước tiên người ta cần xác định hàm mật độ xác suất thực nghiệm đối với X ;

$\hat{f}(e_{1_{i_1}}, e_{2_{i_2}}, \dots, e_{d_{i_d}}) = \frac{1}{n} \sum_{k=1}^n I_{i_1 i_2 \dots i_d}(x_k) = \frac{n_{i_1 i_2 \dots i_d}}{n} = \hat{p}_{i_1 i_2 \dots i_d}$; trong đó $I_{i_1 i_2 \dots i_d}$ là biến chỉ thị, $I_{i_1 i_2 \dots i_d}(x_k) = 1$ nếu $x_{k1} = e_{1_{i_1}}, x_{k2} = e_{2_{i_2}}, \dots, x_{kd} = e_{d_{i_d}}$; 0 nếu ngược lại. Tổng của $I_{i_1 i_2 \dots i_d}$ trên tất cả n điểm của mẫu này là số các xuất hiện $n_{i_1 i_2 \dots i_d}$ của vector

¹ Multiway: nhiều đường.

kí hiệu $(a_{i11}, \dots, a_{did})$. Chia số này cho n sẽ được xác suất quan sát các kí hiệu này. Sử dụng kí pháp $i = (i_1, i_2, \dots, i_d)$ để kí hiệu bộ chỉ mục, người ta có thể viết PMF thực nghiệm kết hợp như ma trận $\hat{P}$ d chiều, kích thước $m_1 * \dots * m_d = \prod_{i=1}^d m_i$. Vậy viết được $\hat{P}(i) = \{\hat{p}_i\}$ đối với tất cả các bộ chỉ mục i , $i_1 = 1..m_1, \dots, i_d = 1..m_d$; $\hat{p}_i = \hat{p}_{i_1 i_2 \dots i_d}$. Bảng ngẫu sinh d chiều là $N = n * \hat{P} = \{n_i\}$ đối với tất cả các bộ chỉ mục i ; $n_i = n_{i_1 i_2 \dots i_d}$. Bảng ngẫu sinh này được tăng thêm

các vector đếm lẻ N_i đối với tất cả d thuộc tính X_i : $N_i = n\hat{p}_i = \begin{pmatrix} n'_1 \\ \dots \\ n'_{m_i} \end{pmatrix}$;

trong đó $\hat{p}_i$ là PMF thực nghiệm đối với X_i .

Kiểm định χ^2 : có thể kiểm định về phụ thuộc d đường giữa d thuộc tính phạm trù, nhờ giả thuyết rỗng H_0 về độc lập d đường. Giả thuyết khác, H_1 , cho rằng chúng không độc lập d đường, tức có phụ thuộc theo đường nào đó. Phân tích ngẫu sinh d chiều cho biết tất cả d thuộc tính là phụ thuộc hay không. Nói chung, cần phân tích ngẫu sinh k đường, để kiểm nghiệm tập con k thuộc tính là phụ thuộc hay không.

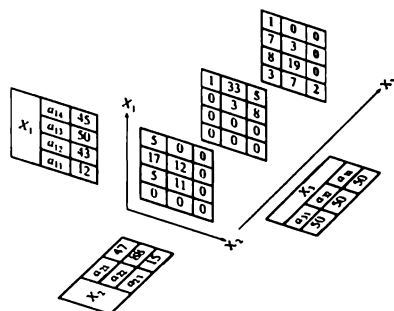
Với giả thuyết rỗng, số kì vọng của các xuất hiện của bộ kí hiệu $(a_{i11}, \dots, a_{did})$ là $e_i = n\hat{p}_i = n \prod_{j=1}^d \hat{p}^j_{i_j} = \frac{n^1_{i_1} n^2_{i_2} \dots n^d_{i_d}}{n^{d-1}}$. Thống kê χ^2 đo sai khác giữa số đếm quan sát n_i và số đếm kì vọng e_i :

$$\chi^2 = \sum_i \frac{(n_i - e_i)^2}{e_i} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \dots \sum_{i_d=1}^{m_d} \frac{(n_{i_1 i_2 \dots i_d} - e_{i_1 i_2 \dots i_d})^2}{e_{i_1 i_2 \dots i_d}}.$$

Thống kê χ^2 theo hàm mật độ χ^2 với q bậc tự do. Đối với bảng ngẫu sinh d đường, có thể tính q do có $\prod_{i=1}^d |dom(X_i)|$ tham số, tức các số đếm, độc lập. Tuy nhiên có thể loại bỏ $\sum_{i=1}^d |dom(X_i)|$ bậc tự do, do vector đếm lẻ theo mỗi chiều X_i cần bằng N_i . Do đã loại bỏ mỗi tham số ấy d lần, cần bổ sung $d - 1$ số đếm tham số tự do, nên tổng số bậc tự do là: $q = \prod_{i=1}^d |dom(X_i)| - \sum_{i=1}^d |dom(X_i)| + (d - 1) = \prod_{i=1}^d m_i - \sum_{i=1}^d m_i + d - 1$.

Để loại bỏ giả thuyết rỗng, cần kiểm tra giá trị p của giá trị quan sát χ^2 so với mức ý nghĩa mong muốn α , khi dùng mật độ χ^2 với q bậc tự do.

Xét bảng ngẫu sinh ba chiều như trong hình, cho biết số đếm quan sát đối với mỗi bộ kí hiệu (a_{1i}, a_{2j}, a_{3k}) đối với ba thuộc tính độ dài lá dài X_1 , độ rộng lá dài X_2 và lớp X_3 .



Hình 4.11. Bảng ngẫu sinh ba đường, với số đếm lẻ theo mỗi chiều

Từ số đếm lẻ đối với X_1 và X_2 trong bảng dữ liệu ngẫu sinh và biết tất cả ba giá trị của X_3 xuất hiện 50 lần, người ta tính được số đếm kì vọng đối với mỗi ô của bảng ngẫu sinh. Theo công thức tính e_i , tính được $e_{(4,1,1)} = \frac{n_{14}^1 * n_{11}^2 * n_{11}^3}{150^2} = \frac{45 * 47 * 50}{150 * 150} = 4.7$. Các số đếm kì vọng là như nhau đối với tất cả ba giá trị của X_3 , như trong bảng sau.

Bảng 4.8. Các số đếm kì vọng 3 đường

		$X_3 (a_{31} / a_{32} / a_{33})$		
		X_2		
		a_{21}	a_{22}	a_{23}
X_1	a_{11}	1.25	2.35	0.40
	a_{12}	4.49	8.41	1.43
	a_{13}	5.22	9.78	1.67
	a_{14}	4.70	8.80	1.50

Giá trị của thống kê χ^2 theo công thức tính χ^2 là $\chi^2 = 231.06$. Tính số bậc tự do, $q = 4 \cdot 3 \cdot 3 - (4 + 3 + 3) + 2 = 28$. Trong hình vẽ, số đếm in đậm là các tham số phụ thuộc. Tất cả các số đếm khác là độc lập. Thực ra, tám ô bất kì chọn ra đều là tham số phụ thuộc.

Đối với mức ý nghĩa $\alpha = 0.01$, giá trị nguy hiểm của phân phối χ^2 là $z = 48.28$. Giá trị quan sát của $\chi^2 = 231.06$ là lớn hơn nhiều so với z , nên càng khó xảy ra giả thuyết rỗng. Người ta kết luận rằng ba thuộc tính không là độc lập 3 đường, mà có phụ thuộc nào đó giữa chúng.

4.2.4. Khoảng cách và góc

Nhờ mô hình hóa các thuộc tính phạm trù như các biến Bernoulli đa biến, có thể tính khoảng cách hay góc giữa hai điểm x_i

và x_j : $x_i = \begin{pmatrix} e_{1,i_1} \\ \dots \\ e_{d,i_d} \end{pmatrix}$, $x_j = \begin{pmatrix} e_{1,j_1} \\ \dots \\ e_{d,j_d} \end{pmatrix}$. Độ đo sai khác về khoảng cách và tính

tương tự ứng với số các giá trị, hay các kí hiệu, khớp và không khớp trên d thuộc tính X_k . Người ta có thể tính số các giá trị khớp s nhờ tích chấm: $s = x_i^T x_j = \sum_{k=1}^d (e_{k,i_k})^T e_{k,j_k}$. Mặt khác, số các giá trị không khớp là $d - s$.

Chuẩn của mỗi điểm là $\|x_i\|^2 = x_i^T x_i = d$.

4.2.4.1. Khoảng cách Oclit

Định nghĩa: Khoảng cách Oclit giữa hai điểm x_i và x_j là $\delta(x_i, x_j) = \|x_i - x_j\| = \sqrt{x_i^T x_i - 2x_i^T x_j + x_j^T x_j} = \sqrt{2(d - s)}$.

Vậy khoảng cách Oclit cực đại giữa hai điểm là $\sqrt{2d}$, khi không có kí hiệu chung giữa chúng, tức $s = 0$.

4.2.4.2. Khoảng cách Hamming

Định nghĩa: Khoảng cách Hamming giữa x_i và x_j được xác định như số các giá trị không khớp nhau: $\delta_H(x_i, x_j) = d - s = \delta(x_i, x_j)^2 / 2$.

Khoảng cách Hamming tương đương một nửa khoảng cách Oclit bình phương.

4.2.4.3. Tương tự cosin

Hàm cosin của góc giữa x_i và x_j là $\cos\theta = x_i^T x_j / (\|x_i\| * \|x_j\|) = s / d$.

4.2.4.4. Hệ số Jaccard

Hệ số Jaccard là công cụ thông dụng để đo sự tương tự giữa hai điểm phạm trù.

Định nghĩa: Hệ số Jaccard được xác định như tỉ số giữa số các giá trị khớp và số các giá trị khác nhau xuất hiện trong cả x_i và x_j , trên d thuộc tính.

Nó được xác định $J(x_i, x_j) = s / (2(d - s) + s) = s / (2d - s)$. Quan sát thấy khi hai điểm không khớp trên k chiều, người ta bổ sung số 2 vào đếm số khác nhau; ngược lại nếu chúng khớp, số các kí hiệu khớp tăng 1. Do có $d - s$ không khớp và s khớp, nên số các kí hiệu khác nhau là $2(d - s) + s$.

Xét dữ liệu phạm trù ba chiều trong thí dụ trước. Điểm (ngắn, vừa, nhiều màu) và (rất ngắn, vừa, mượt) được mô hình hóa như vector trong R^{10} .

$$x_1 = \begin{pmatrix} e_{12} \\ e_{22} \\ e_{31} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \\ 1 & 0 & 0 & \end{pmatrix}, \quad x_2 = \begin{pmatrix} e_{11} \\ e_{22} \\ e_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \\ 0 & 1 & 0 & \end{pmatrix}$$

Số các kí hiệu khớp là $s = x_1^T x_2 = (e_{12})^T e_{11} + (e_{22})^T e_{22} + (e_{31})^T e_{32} =$

$$(0 \ 1 \ 0 \ 0) \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + (0 \ 1 \ 0) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + (1 \ 0 \ 0) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = 0 + 1 + 0 = 1. \text{ Đã có } d = 3$$

và s , có thể tính khoảng cách $\delta(x_1, x_2) = \sqrt{2(d-s)} = 2$; $\delta H(x_1, x_2) = d - s = 2$. Độ tương tự cosin và Jaccard là $\cos\theta = s/d = 0.333$; $J(x_1, x_2) = s/(2d - s) = 0.2$.

4.2.5. Phân biệt

Định nghĩa: Phân biệt, hay xếp ngăn¹, là việc chuyển các thuộc tính số sang thuộc tính phạm trù.

Người ta sử dụng khả năng phân biệt trong các phương pháp khai phá dữ liệu không quản lí các dữ liệu số. Nó cho phép giảm số các giá trị của thuộc tính, đặc biệt khi có ồn trong số đo số; cho phép bỏ qua các sai khác không rõ ràng và sai nhỏ trong các giá trị.

Về hình thức, cho thuộc tính số X và mẫu ngẫu nhiên $\{x_i\}$, $i = 1..n$ kích thước n từ X , nhiệm vụ phân biệt sẽ chia khoảng giá trị của X thành k khoảng kế tiếp, được gọi là *ngăn²*, thông qua việc phát hiện các giá trị biên $v_1, v_2, \dots, v_{k-1}$, tạo nên k khoảng $[x_{\min}, v_1]$, $(v_1, v_2], \dots, (v_{k-1}, v_{\max}]$; trong đó $x_{\min} = \min\{x_i\}$, $x_{\max} = \max\{x_i\} \forall i$.

Kết quả là k khoảng, hay k ngăn, phủ toàn bộ khoảng của X , được ánh xạ sang các giá trị kí hiệu của thuộc tính X phạm trù mới.

4.2.5.1. Khoảng rộng bằng nhau

Tiếp cận các ngăn đơn giản nhất là phân đoạn khoảng X thành k khoảng có *độ rộng bằng nhau*. Độ rộng của khoảng là $w = (x_{\max} - x_{\min})/k$. Biên của khoảng thứ i là $v_i = x_{\min} + i \cdot w$.

4.2.5.2. Khoảng có tần suất như nhau

Tiếp cận *tần suất như nhau* sẽ được các khoảng có số các điểm bằng nhau. Khó tính tần suất như nhau do có các giá trị lặp. Các khoảng này được tính từ các điểm *phân vị³* thực nghiệm hay hàm phân

¹ Discretization: phân biệt; binning: xếp ngăn.

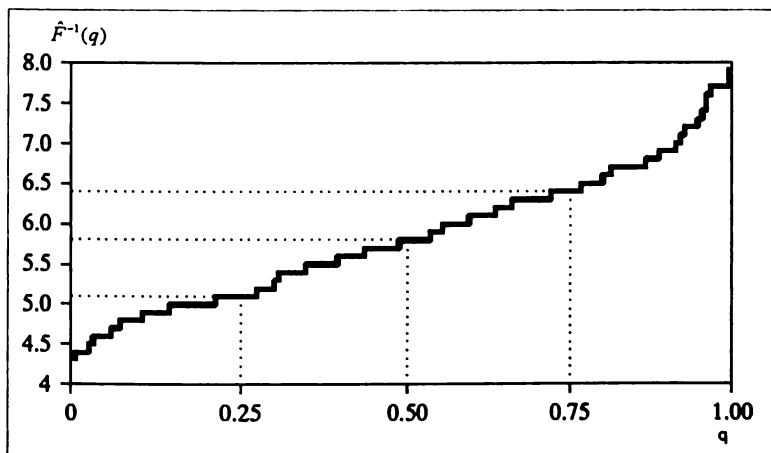
² Bin: ngăn.

³ Quantile: điểm phân vị.

phối tích lũy ngược $\hat{F}^{-1}(q)$ đối với X . $\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$, với $q \in [0; 1]$. Cụ thể người ta cần mỗi khoảng chứa $1/k$ của mật độ xác suất; vậy các biên của khoảng là $v_i = \hat{F}^{-1}(i/k)$.

Xét thuộc tính độ dài lá đài trong bộ dữ liệu hoa diên vĩ. Các giá trị $x_{\min} = 4.3$, $x_{\max} = 7.9$. Người ta phân biệt các giá trị thuộc tính vào k ngăn, $k = 4$, theo độ rộng đều nhau. Độ rộng $w = (7.9 - 4.3)/4 = 0.9$. Do vậy $v_1 = 4.3 + 0.9 = 5.2$; $v_2 = 4.3 + 2 \cdot 0.9 = 6.1$; $v_3 = 4.3 + 3 \cdot 0.9 = 7.0$. Các ngăn này đã thể hiện trong bảng đầu chương, cho thấy số các điểm n_i trong mỗi ngăn, và không cân bằng giữa các ngăn.

Phân biệt theo tần suất, người ta xét hàm phân phối tích lũy CDF ngược đối với độ dài lá đài, như trong hình vẽ. Với $I = 4$, biên của mỗi giá trị phân v_i , ứng với đường gạch gạch trong hình, là $v_1 = \hat{F}^{-1}(0.25) = 5.1$; $v_2 = \hat{F}^{-1}(0.50) = 5.8$ và $v_3 = \hat{F}^{-1}(0.75) = 6.4$.



Hình 4.12. Hàm CDF ngược thực nghiệm đối với độ dài lá đài

Các khoảng thu được thể hiện trong bảng sau. Người ta thấy dù độ rộng khoảng thay đổi, chúng chứa số điểm khá cân đối. Không thể đạt được số điểm bằng nhau, do có chín điểm có giá trị 5.1, bảy điểm có giá trị 5.8.

Bảng 4.9. Phân biệt đều theo tần suất đối với độ dài lá dài

Ngăn	Độ rộng	Đếm
[4.3, 5.1]	0.8	$n_1 = 41$
(5.1, 5.8]	0.7	$n_2 = 39$
(5.8, 6.4]	0.6	$n_3 = 35$
(6.4, 7.9]	1.5	$n_4 = 35$

4.3. Dữ liệu đồ thị

Khung truyền thống về phân tích dữ liệu thường đặt giả thiết về tính độc lập của các dữ liệu. Tuy nhiên, thực sự các dữ liệu có mối liên kết. Chúng tạo nên đồ thị, mà mỗi nút, cung mang thuộc tính số hay phạm trù. Nhiều ứng dụng cần đến dữ liệu đồ thị, chẳng hạn như WWW và các siêu liên kết, mạng xã hội và các dữ liệu đa phương tiện, mạng sinh học với tương tác hữu cơ hay mạng các gen, mạng trích dẫn đối với tác phẩm khoa học...

4.3.1. Khái niệm về đồ thị

4.3.1.1. Đồ thị

Về hình thức, đồ thị là cấu trúc toán học, gồm tập hữu hạn không rỗng các đỉnh và tập các cạnh.

Định nghĩa: Đồ thị $G = (V, E)$ là cấu trúc toán học, V là tập hữu hạn khác rỗng, gọi là đỉnh hay nút; $E \subset V * V$ là các cặp không thứ tự của các đỉnh, gọi là cạnh.

Cạnh nối nút đến chính nút ấy được gọi là *khuyên*¹. Đồ thị không có hướng, không lặp, được gọi là *đồ thị đơn giản*. Nếu không nêu rõ, đồ thị được hiểu là đồ thị đơn giản. Cạnh $e = (v_i, v_j)$ được gọi là *nối với*² v_i, v_j ; hai nút này được gọi là *kề nhau*, là *láng giềng* của nhau. Số nút trong đồ thị là $|V| = n$, được gọi là *cấp*³ của đồ thị; số cạnh $|E| = m$ được gọi là *kích thước* của đồ thị.

¹ Loop: khuyên.

² Incident with: nối với.

³ Order, size: cấp, kích thước (của đồ thị).

Định nghĩa: Đồ thị có hướng¹ là đồ thị mà tập cạnh E gồm các cặp có thứ tự của đỉnh.

Cạnh có hướng (v_i, v_j) được gọi là cung², được xác định như từ v_i đến v_j ; v_i là đuôi, v_j là đầu của cung này. Đồ thị có trọng số là đồ thị với trọng số w_i đối với mỗi cạnh $(v_i, v_j) \in E$. Mọi đồ thị đều có sẵn trọng số 1.

4.3.1.2. Đồ thị con

Định nghĩa: Đồ thị $H = (VH, EH)$ là đồ thị con của $G = (V, E)$ nếu $VH \subseteq V, EH \subseteq E$.

Đồ thị H là con của G có nghĩa G là đồ thị cha³ của H . Lấy $V' \subseteq V$, các đồ thị con dẫn xuất⁴ $G' = (V', E')$ sẽ mang các cạnh của G , nối các đỉnh của V' . Hai nút là kề nhau trong G' chỉ khi là kề nhau trong G . Đồ thị là đầy đủ nếu có cạnh nối tất cả các cặp đỉnh.

4.3.1.3. Bậc

Định nghĩa: Bậc của nút $v_i \in V$ là số các cạnh nối với nó, kí hiệu là $d(v_i)$, hay d_i .

Chuỗi bậc⁵ của đồ thị là danh sách của các bậc của các nút được sắp xếp theo thứ tự không tăng. Lấy N_k là số các đỉnh có bậc k . Phân phối tần suất bậc của đồ thị là $(N_0, N_1, \dots, N_t)$, trong đó t là bậc cực đại đối với nút trong G . Phân phối bậc của đồ thị cho biết hàm mật độ xác suất f đối với X , tức $(f(0), f(1), \dots, f(t))$, trong đó $f(k) = P(X = k) = N_k / n$ là xác suất để nút có bậc k , như số của các nút N_k với bậc k chia cho tổng số nút n . Trong phân tích đồ thị, người ta giả thiết đồ thị đầu vào biểu diễn lực lượng, nên người ta viết f thay cho $\hat{f}$ đối với phân phối xác suất.

¹ Direct graph, digraph: đồ thị có hướng.

² Arc: cung.

³ Supergraph: đồ thị cha.

⁴ Induced graph: đồ thị dẫn xuất.

⁵ Degree sequence: chuỗi bậc.

Đối với đồ thị có hướng, *bậc đến*¹ của nút v_i , kí hiệu $id(v_i)$, là số các cạnh lấy v_i là đầu, tức số các cạnh vào v_i . *Bậc đến* của v_i , kí hiệu $od(v_i)$, là số các cạnh lấy v_i làm cuối, tức số các cạnh đi ra từ v_i .

4.3.1.4. Đường và khoảng cách

*Định nghĩa: Đường*² trong đồ thị G giữa các nút x và y là xâu có thứ tự các đỉnh, bắt đầu từ x , kết thúc tại y .

Đường (v_i) có $x = v_0, v_1, \dots, v_t = y$, có cạnh nối cặp đỉnh liên tiếp, tức $(v_{i-1}, v_i) \in E$. Chiều dài của đường, t , được đo theo thuật ngữ *bước nhảy*³, tức số các cạnh dọc theo đường. Trong đường, không hạn chế về số lần xuất hiện của đỉnh; nên cả đỉnh và cạnh có thể lặp lại. Đường có đỉnh đầu và cuối trùng nhau gọi là *đóng*. *Đường đi đơn*⁴ là đường có các đỉnh phân biệt, tức qua nhiều nhất là một lần. *Vạch* là đường có các cạnh phân biệt. Đường đơn đóng với chiều dài $t \geq 3$ được gọi là chu trình. Đường đơn có chiều dài ngắn nhất giữa nút x, y được gọi là đường ngắn nhất, và chiều dài này được gọi là độ dài giữa x và y , kí hiệu $d(x, y)$.

Định nghĩa: Độ dài giữa nút x và y , kí hiệu $d(x, y)$ là đường đơn ngắn nhất giữa x, y .

4.3.1.5. Liên thông

*Định nghĩa: Hai nút v_i và v_j là liên thông*⁵ nếu có đường đơn giữa chúng.

Thành phần liên thông, hay *thành phần*, của đồ thị là đồ thị con liên thông lớn nhất. Nếu đồ thị chỉ có một thành phần, nó là liên thông; ngược lại, nó là không liên thông, do theo định nghĩa không có đường nối giữa hai thành phần khác nhau.

¹ Indegree, outdegree: bậc đến, bậc đi.

² Walk: đường.

³ Hops: bước nhảy.

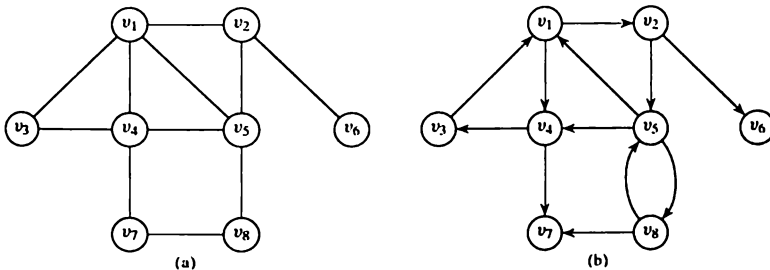
⁴ Path: đường đơn; trail: vạch.

⁵ Connected: liên thông.

Định nghĩa: Đồ thị liên thông là đồ thị có đường giữa tất cả các cặp đỉnh.

Đối với đồ thị có hướng, đồ thị là *liên thông mạnh* nếu có đường, có hướng, giữa tất cả các cặp có thứ tự của các đỉnh. Người ta nói đồ thị là *liên thông yếu*, nếu có đường giữa cặp nút theo các cạnh, như đồ thị không có hướng.

Hình sau cho thấy đồ thị có $|V| = 8$ đỉnh, $|E| = 11$ cạnh. Do $(v_1, v_5) \in E$, có thể nói rằng v_1, v_5 kề nhau. Bậc của v_1 là $d(v_1) = d_1 = 4$. Xâu các bậc của đồ thị là $(4, 4, 4, 3, 2, 2, 2, 1)$ nên phân phối tần suất bậc của đồ thị này là $(N_0, N_1, N_2, N_3, N_4) = (0, 1, 3, 1, 3)$. Người ta có $N_0 = 0$ do không có đỉnh cô lập, và $N_4 = 3$ do có ba nút v_1, v_4 và v_5 , nên bậc là $k = 4$; tương tự có các số khác. Phân phối bậc là $(f(0), f(1), f(2), f(3), f(4)) = (0, 0.125, 0.375, 0.125, 0.375)$. Xâu các đỉnh $(v_3, v_1, v_2, v_5, v_1, v_2, v_6)$ là đường với chiều dài 6, giữa v_3 và v_6 . Người ta thấy đỉnh v_1 và v_2 được thăm quá một lần. Dãy đỉnh $(v_3, v_4, v_7, v_8, v_5, v_2, v_6)$ là đường chiều dài 6 giữa v_3 và v_6 . Tuy nhiên đây không là đường ngắn nhất giữa chúng; ngắn là (v_3, v_1, v_2, v_6) có chiều dài 3. Vậy khoảng cách giữa chúng $d(v_3, v_6) = 3$.



Hình 4.13. Đồ thị có hướng và đồ thị vô hướng

Hình bên cạnh cho thấy đồ thị có hướng với 8 đỉnh và 12 cạnh. Thấy rằng cạnh (v_5, v_8) khác với (v_8, v_5) . Bậc vào của v_7 là $id(v_7) = 2$; bậc ra $od(v_7) = 0$. Vậy không có đường, có hướng, từ v_7 đến bất kì đỉnh nào.

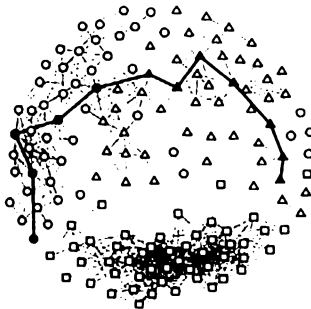
4.3.1.6. Ma trận kề

Đồ thị $G = (V, E)$ với $|V| = n$ đỉnh có thể được biểu diễn dưới dạng ma trận kề nhị phân đối xứng $n \times n$, A , với $A(i, j) = 1$ nếu v_i kề v_j ; 0 nếu ngược lại. Nếu đồ thị là có hướng, ma trận kề A không đối xứng, do $(v_i, v_j) \in E$ không ngầm định $(v_j, v_i) \in E$.

Nếu đồ thị có trọng số, thì người ta có ma trận kề có trọng số, A , với $A(i, j) = w_{ij}$ nếu v_i kề v_j ; 0 nếu ngược lại; trong đó w_{ij} là trọng số của cạnh (v_i, v_j) . Ma trận kề có trọng số luôn có thể được chuyển thành ma trận nhị phân, nếu sử dụng ngưỡng τ , tức $A(i, j) = 1$ nếu $w_{ij} > \tau$; 0 nếu ngược lại.

4.3.1.7. Các đồ thị từ ma trận dữ liệu

Có thể chuyển nhiều bộ dữ liệu không theo dạng đồ thị sang dạng đồ thị. Gọi $D = \{x_i\}$, $i = 1..n$, với $x_i \in \mathbb{R}^d$, là bộ dữ liệu có n điểm trong không gian d chiều. Người ta có thể xác định đồ thị có trọng số $G = (V, E)$, mà nút ứng với mỗi điểm của D , và có cung giữa cặp điểm, với trọng số $w_{ij} = \text{sim}(x_i, x_j)$; trong đó $\text{sim}(x_i, x_j)$ là độ tương tự giữa điểm x_i và x_j . Độ tương tự có thể xác định như nghịch đảo của khoảng cách Oclit giữa các điểm: $w_{ij} = \text{sim}(x_i, x_j) = \frac{\|x_i - x_j\|^2}{2\sigma^2}$, mà σ là tham số trải, tương đương với độ lệch trong hàm mật độ chuẩn. Việc chuyển này hạn chế hàm $\text{sim}()$ nằm trong $[0, 1]$. Có thể chọn ngưỡng τ và chuyển ma trận kề có trọng số về ma trận nhị phân.



Hình 4.14. Đồ thị về tính tương tự của bộ dữ liệu hoa diên vĩ

Hình vẽ cho thấy đồ thị tương tự đối với bộ dữ liệu hoa diên vĩ. Độ tương tự theo cặp giữa các cặp khác nhau của các điểm được tính theo công thức trên, với $\sigma = 1/\sqrt{2}$. Độ tương tự trung bình giữa các điểm là 0.197, và độ lệch chuẩn là 0.290. Ma trận kề nhị phân, với ngưỡng $\tau = 0.777$ cho kết quả trong cạnh giữa các điểm có độ tương tự cao hơn hai lần độ lệch chuẩn so với trung bình. Đồ thị hoa diên vĩ có 150 nút và 753 cạnh. Các nút trong hình vẽ phân lớp phạm trù theo lớp của chúng. Dấu tròn cho biết lớp *nhiều màu*, tam giác có hoa, hình vuông *mượt*. Đồ thị có hai thành phần lớn; một cái chỉ gồm các nút với nhãn *mượt*.

4.3.2. Các thuộc tính hình dạng

Phần này đề cập thuộc tính cấu trúc, dựa trên cạnh của đồ thị. Các thuộc tính này là *cục bộ* nếu chúng chỉ áp lên một nút, hay một cạnh, và là *toàn thể* nếu dùng cho toàn bộ đồ thị.

4.3.2.1. Bậc

Định nghĩa tổng quát đối với bậc của nút v_i , khi đồ thị có trọng số là $d_i = \sum_j A(i, j)$. Bậc là thuộc tính cục bộ. Một thuộc tính tổng quát đơn giản nhất là *bậc trung bình*: $\mu_d = \sum_i d_i / n$. Định nghĩa có thể phát biểu cho đồ thị có hướng, có trọng số. Chẳng hạn có thể xét bậc vào, bậc ra: $id(v_i) = \sum_j A(j, i)$ và $od(v_i) = \sum_j A(i, j)$. Trung bình của bậc ra, vào như nhau.

4.3.2.2. Độ dài đường trung bình

Định nghĩa: Độ dài đường trung bình, hay độ dài đường đặc trưng, của đồ thị liên thông là

$$\mu_L = \sum_i \sum_{j, i \neq j} d(v_i, v_j) / \binom{n}{2} = \frac{2}{n(n-1)} \sum_i \sum_{j, i \neq j} d(v_i, v_j).$$

Trong định nghĩa trên, n là số nút trong đồ thị, $d(v_i, v_j)$ là khoảng cách giữa v_i và v_j . Đối với đồ thị liên thông, trung bình

tính trên tất cả các cặp đỉnh có thứ tự: $\mu_L = \frac{1}{n(n-1)} \sum_i \sum_j d(v_i, v_j)$.

Đối với đồ thị không liên thông, trung bình tính chỉ trên các cặp đỉnh nối nhau.

4.3.2.3. Tính lệch tâm

Định nghĩa: *Lệch tâm*¹ của nút v_i là khoảng cách cực đại từ v_i đến nút bất kì trong đồ thị.

Lệch tâm $e(v_i) = \max \{d(v_i, v_j)\}$ theo j . Nếu đồ thị là không liên thông, lệch tâm được tính chỉ trên các cặp đỉnh với khoảng cách hữu hạn, tức chỉ với đỉnh nối với đồ thị.

4.3.2.4. Bán kính và đường kính

Định nghĩa: *Bán kính* của đồ thị liên thông, kí hiệu $r(G)$, là lệch tâm tối thiểu của bất kì nút nào trong đồ thị này. *Đường kính*, kí hiệu $d(G)$, là lệch tâm cực đại của bất kì đỉnh nào trong đồ thị.

$$r(G) = \min_i \{e(v_i)\} = \min_i \{\max_j \{d(v_i, v_j)\}\}$$

$$d(G) = \max_i \{e(v_i)\} = \max_{i,j} \{d(v_i, v_j)\}$$

Đối với đồ thị không liên thông, đường kính là lệch tâm cực đại trên tất cả các thành phần liên thông của đồ thị này. Đường kính của đồ thị G nhạy với dữ liệu lẻ. Khái niệm bền vững hơn là *đường kính hữu hiệu*², xác định như số tối thiểu các bước nhảy để tỉ lệ lớn, thường là 90%, các cặp nút liên thông đến được với nhau.

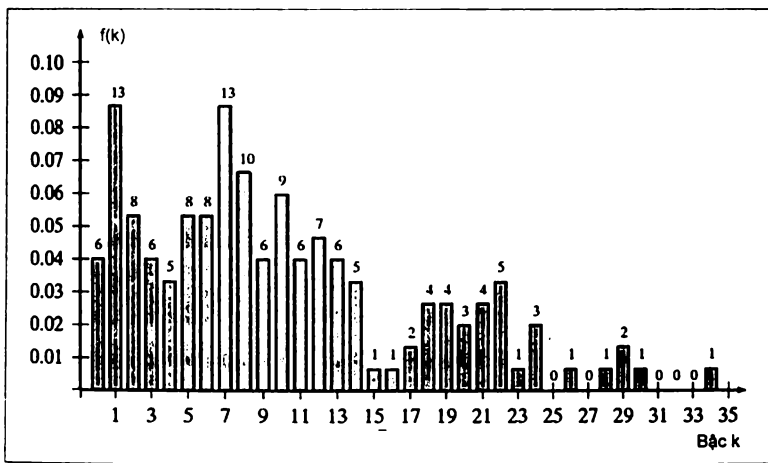
Định nghĩa: *Đường kính hữu hiệu* là giá trị k nhỏ nhất mà $H(k) \geq 0.9 * H(d(G))$.

$H(k)$ là số các cặp nút đến được nhau, sau k bước nhảy, hay ít hơn.

¹ Eccentricity: tính lệch tâm.

² Effective diameter: đường kính hữu hiệu.

Xét đồ thị trong thí dụ trước, lệch tâm của nút v_4 là $e(v_4) = 3$, do nút xa nhất từ nó là v_6 , và $d(v_4, v_6) = 3$. Bán kính của đồ thị là $r(G) = 2$; cả v_1 và v_5 có lệch tâm nhỏ nhất, là 2. Đường kính của đồ thị là $d(G) = 4$, do khoảng cách lớn nhất trên các cặp là $d(v_6, v_7) = 4$.

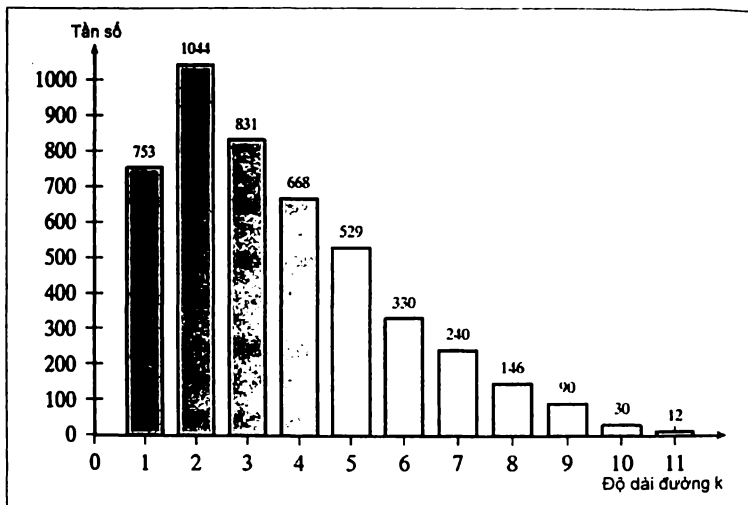


Hình 4.15. Phân phối bậc của đồ thị hoa diên vĩ

Đường kính của đồ thị hoa diên vĩ là $d(G) = 11$, ứng với đường đậm nối các nút xám. Phân phối bậc đối với đồ thị hoa diên vĩ như trong hình sau.

Số trên đỉnh thanh đứng cho biết tần suất. Chẳng hạn có 13 nút có bậc 7, ứng với xác suất $f(7) = 13/150 = 0.0867$.

Có 1044 cặp nút có khoảng cách 2 bước nhảy. Với $n = 150$ nút, có $\binom{n}{2} = 11175$ cặp. Có 6502 cặp này là không liên thông, và tất cả có 4673 cặp nối nhau. Có $4175/4673 = 0.89$ nối nhau bằng 6 bước nhảy, và $4415/4673 = 0.94$ nối nhau theo 7 bước. Vậy người ta có thể định được đường kính hiệu quả là 7. Chiều dài trung bình của đường là 3.58.



Hình 4.16. Biểu đồ chiều dài đường, trên đồ thị hoa diên vĩ

4.3.2.5. Hệ số phân cụm

Định nghĩa: Hệ số phân cụm của nút v_i là độ đo mật độ các cạnh láng giềng của v_i .

Lấy $G_i = (V_i, E_i)$ là đồ thị con rút ra nhờ lấy láng giềng của đỉnh v_i ; $v_i \notin V_i$, $|V_i| = n_i$ là số láng giềng của v_i , $|E_i| = m_i$ là số cạnh trong các láng giềng của v_i . Hệ số phân cụm của v_i là $C(v_i)$, là tỉ số giữa số các cạnh trong G_i và số cực đại các cạnh trong G_i , tức
$$C(v_i) = \frac{m_i}{\binom{n_i}{2}} = \frac{2m_i}{n_i(n_i - 1)}.$$
 Hệ số phân cụm cho dấu hiệu của riêng về

láng giềng của một nút, do mẫu số ứng với trường hợp khi G_i là đồ thị con đầy đủ.

Hệ số phân cụm của đồ thị G đơn giản là hệ số phân cụm trung bình trên toàn bộ các nút, tức $C = \frac{1}{n} \sum_i C(v_i)$. Do $C(v_i)$ xác định chỉ với các nút có bậc $d(v_i) \geq 2$, người ta có thể cho $C(v_i) = 0$ đối với các

nút có bậc nhỏ hơn 2. Hoặc có thể giả thiết việc lấy trung bình chỉ với các nút có $d(v_i) \geq 2$.

Hệ số phân cụm $C(v_i)$ của một nút liên quan chặt chẽ với khái niệm về mối quan hệ bắc cầu trong đồ thị hay mạng. Tức nếu có cạnh giữa v_i và v_j , và cạnh khác giữa v_i và v_k , thì v_j và v_k nối nhau. Xác định đồ thị con gồm cạnh (v_i, v_j) và (v_i, v_k) như bộ ba liên thông, với tâm là v_i . Bộ ba liên thông, tâm v_i , có (v_j, v_k) , gọi là tam giác, hay đồ thị con kích thước 3. Hệ số phân cụm của nút v_i được thể hiện: $C(v_i)$ là tỉ số giữa số các tam giác chứa v_i và số các bộ ba liên thông với tâm v_i . Số các bộ ba liên thông có tâm v_i là $\binom{d_i}{2} = \frac{n_i(n_i - 1)}{2}$; trong đó $d_i = n_i$ là số các láng giềng của v_i .

Tổng quát hóa khái niệm vừa nêu cho đồ thị G , người ta có tính bắc cầu của đồ thị này là $T(G)$, là tỉ số giữa ba lần số tam giác trong G và số các bộ ba liên thông trong G . Nhân tử 3 trên tử số là do mỗi tam giác cho ba bộ ba liên thông, với tâm là ba đỉnh.

4.3.2.6. Hiệu quả

Định nghĩa: Hiệu quả của cặp nút v_i, v_j là $1/d(v_i, v_j)$.

Nếu v_i không nối với v_j , thì $d(v_i, v_j) = \infty$, nên hiệu quả là 0. Khoảng cách giữa các nút càng nhỏ, thì liên lạc giữa chúng càng hiệu quả.

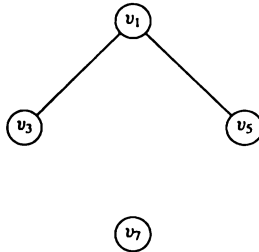
Định nghĩa: Hiệu quả của đồ thị G là hiệu quả trung bình của tất cả các cặp nút, dù chúng có nối nhau hay không.

Hiệu quả của đồ thị G là $\frac{2}{n(n-1)} \sum_i \sum_{j>i} \frac{1}{d(v_i, v_j)}$. Hiệu quả

cực đại là 1, ứng với đồ thị đầy đủ.

Hiệu quả cục bộ đối với nút v_i được xác định như hiệu quả của đồ thị con G_i , sinh ra nhờ các láng giềng của v_i . Do $v_i \notin G_i$, hiệu quả cục bộ là chỉ thị về dung thứ lỗi cục bộ, tức liên lạc giữa các láng giềng của v_i là hiệu quả ra sao khi v_i bị loại khỏi đồ thị này.

Đối với đồ thị đã nêu, xét nút v_4 . Đồ thị láng giềng của nó như hình sau. Hệ số cụm của nút v_4 là $C(v_4) = \frac{2}{\binom{4}{2}} = \frac{2}{6} = 0.33$.



Hình 4.17. Đồ thị con G_4 sinh từ nút v_4

Hệ số phân cụm đối với toàn đồ thị, tức trên tất cả các nút, là $C(G) = 1/8 * (1/2 + 1/3 + 1 = 1/3 + 1/3 + 0 + 0 + 0) = 2.5/8 = 0.3125$. Hiệu quả cục bộ của v_4 là:

$$\begin{aligned}
 & \frac{2}{4*3} \left(\frac{1}{d(v_1, v_3)} + \frac{1}{d(v_1, v_5)} + \frac{1}{d(v_1, v_7)} + \right. \\
 & \left. + \frac{1}{d(v_3, v_5)} + \frac{1}{d(v_3, v_7)} + \frac{1}{d(v_5, v_7)} \right) \\
 & = \frac{1}{6} (1 + 1 + 0 + 0.5 + 0 + 0) = 0.417
 \end{aligned}$$

4.3.3. Phân tích hướng tâm

Khái niệm *hướng tâm*¹ được dùng để phân hạng các đỉnh của đồ thị, theo thuật ngữ về hướng tâm ra sao, hay chúng quan trọng thế nào. Tính hướng tâm có thể hình thức hóa như hàm $c: V \rightarrow R$, cho biết *cấp*² tổng cộng trên V . Người ta nói v_i ít nhất cũng hướng tâm như v_j nếu $C(v_i) \geq C(v_j)$.

¹ Centrality: hướng tâm.

² Order: cấp.

4.3.3.1. Tính hướng tâm cơ bản

- *Hướng tâm về bậc*: khái niệm đơn giản nhất về hướng tâm là bậc di của đỉnh v_i ; bậc càng cao thì đỉnh càng quan trọng, hay càng ở tâm. Đối với đồ thị có hướng, có thể xét hướng tâm bậc trong hay bậc ngoài.
- *Hướng tâm lệch*: theo khái niệm này, một nút là ít hướng tâm lệch thì nó sẽ càng hướng tâm. Hướng tâm lệch được xác định theo $C(v_i) = 1/e(v_i) = 1/\max_j\{d(v_i, v_j)\}$. Một nút v_i ít lệch, tức lệch bằng bán kính đồ thị, $e(v_i) = r(G)$, thì nút này là *nút giữa*; một nút lệch nhất, tức lệch bằng đường kính, $e(v_i) = d(G)$, được gọi là nút *ngoại biên*. Hướng tâm lệch liên quan đến bài toán về *định vị tiện nghi*, tức chọn vị trí tối ưu cho tài nguyên hay tiện nghi. Nút giữa tối thiểu hóa khoảng cách tối đa đến nút bất kì trong mạng, nên nút giữa nhất sẽ có vị trí lí tưởng.
- *Hướng tâm đóng*: trong khi hướng tâm lệch sử dụng cực đại khoảng cách từ nút đã cho, *hướng tâm đóng*¹ sử dụng tổng tất cả các khoảng cách để phân hạng một nút ở giữa ra sao.

Hướng tâm đóng là $c(v_i) = \frac{1}{\sum_j d(v_i, v_j)}$. Nút v_i với tổng

khoảng cách nhỏ nhất được gọi là *nút trung vị*². Tính hướng tâm đóng tối ưu hàm khách quan khác nhau đối với bài toán định vị tiện nghi. Nó cố gắng tối thiểu hóa khoảng cách tổng trên tất cả các nút khác. Nút trung vị có hướng tâm đóng lớn nhất, cũng là lời giải tối ưu.

- *Hướng tâm giữa*: đối với đỉnh v_i đã cho, *hướng tâm giữa* đo các đường ngắn nhất nhiều ra sao, giữa tất cả các cặp đỉnh có v_i . Nó cho biết vai trò giám sát ở giữa của v_i , đối với các cặp nút. Lấy n_{jk} là số các đường ngắn nhất giữa đỉnh v_i và v_k , và kí hiệu $n_{jk}(v_i)$ là số các đường như vậy có chứa v_i . Phần đường qua v_i là $g_{jk}(v_i) = n_{jk}(v_i)/n_{jk}$. Nếu hai đỉnh v_j và v_k

¹ Closeness centrality: hướng tâm đóng.

² Median node: nút trung vị.

không nối nhau, người ta lấy $g_{jk}(v_i) = 0$. Hướng tâm giữa đối với nút v_i được xác định theo:

$$c(v_i) = \sum_{j \neq i} \sum_{k \neq i, k \neq j} g_{jk}(v_i) = \sum_{j \neq i} \sum_{k \neq i, k \neq j} \frac{n_{jk}(v_i)}{n_{jk}}.$$

Xét đồ thị trước đây. Các giá trị đối với các số đo hướng tâm của các nút khác nhau thể hiện trong bảng sau.

Bảng 4.10. Các giá trị hướng tâm

Hướng tâm	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Bậc	4	3	2	4	4	1	2	2
Lệch	0.5	0.33	0.33	0.33	0.5	0.25	0.25	0.33
$e(v_i)$	2	3	3	3	2	4	4	3
Đóng	0.100	0.083	0.071	0.091	0.100	0.056	0.067	0.071
$\sum_j d(v_i, v_j)$	10	12	14	11	10	18	15	14
Giữa	4.5	6	0	5	6.5	0	0.83	1.17

Theo hướng tâm theo bậc, nút v_1 , v_4 và v_5 là hướng tâm nhất. Hướng tâm lệch là cao nhất đối với các nút giữa trong đồ thị, là v_1 và v_5 . Số đo này thấp nhất đối với các nút ngoại biên, là v_6 và v_7 .

Các nút v_1 và v_5 có giá trị hướng tâm đóng cao nhất. Theo thuật ngữ về giữa, đỉnh v_5 là ở giữa nhất, với giá trị 6.5. Người ta có thể tính giá trị này khi chỉ xét các cặp nút v_i và v_k có ít nhất một đường ngắn nhất qua v_5 , vì chỉ các cặp nút này có $g_{jk}(v_5) > 0$.

$$C(v_5) = g_{18}(v_5) + g_{24}(v_5) + g_{27}(v_5) + g_{28}(v_5) + g_{38}(v_5) + g_{46}(v_5) + g_{48}(v_5) + g_{67}(v_5) + g_{68}(v_5) = 1 + 1/2 + 2/3 + 1 + 2/3 + 1/2 + 1/2 + 2/3 + 1 = 6.5.$$

4.3.3.2. Hướng tâm Web

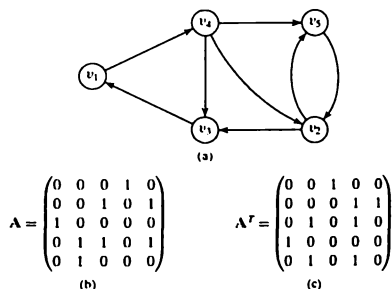
Xét đồ thị có hướng, đặc biệt trong ngữ cảnh của Web. Chẳng hạn các tư liệu siêu văn bản có các kết nối có hướng từ tư liệu này

đến tư liệu khác, hay mạng trích dẫn các bài báo khoa học. Khái niệm hướng tâm đặc biệt phù hợp với các đồ thị trong Web.

Uy tín: khái niệm *uy tín*¹, hay *hướng tâm vector riêng*, của một nút trong đồ thị có hướng, liên quan đến độ đo về độ quan trọng hay xếp hạng của một nút. Về trực quan, càng nhiều liên kết đến nút đã biết, uy tín nút đó càng cao. Tuy nhiên, uy tín không phụ thuộc đơn giản vào bậc trong, mà còn phụ thuộc vào uy tín của nút trỏ đến nó.

Lấy $G = (V, E)$ là đồ thị có hướng, với $|V| = n$. Ma trận kề của G là ma trận bất đối xứng $n \times n$, với $A(u, v) = 1$ nếu $(u, v) \in E$, 0 nếu ngược lại. Cho $p(u)$ là số thực dương, gọi là *điểm uy tín* của nút u . Uy tín của nút phụ thuộc vào uy tín của các nút khác, trỏ đến nó, nên người ta có điểm uy tín của nút v là: $p(v) = \sum_u A(u, v) \cdot p(u) = \sum_u A^T(v, u) \cdot p(u)$.

Chẳng hạn với hình vẽ sau, uy tín của v_5 phụ thuộc vào uy tín của v_2 và v_4 .



Hình 4.18. Đồ thị, ma trận kề và ma trận chuyển vị của nó

Duyệt các nút, người ta có thể biểu diễn đệ quy điểm uy tín theo $p' = A^T p$, trong đó p là vector cột ứng với điểm uy tín đối với mỗi đỉnh. Nếu p_{k-1} là điểm uy tín theo tất cả các nút tại lần lặp $k-1$, điểm cập nhật tại lần lặp k là $p_k = A^T p_{k-1} = A^T (A^T p_{k-2}) = \dots = (A^T)^k p_0$; trong đó p_0 là vector uy tín khởi đầu. Rõ ràng là vector p_k

¹ Prestige: uy tín.

hội tụ đến vector riêng trội¹ của A^T lũy thừa k . Vector trội của A^T và giá trị riêng tương ứng có thể được tính theo tiếp cận *lặp lũy thừa* như trong thuật toán sau.

Thuật toán: phương pháp lặp lũy thừa, vector riêng trội.

Lặp lũy thừa (A, ϵ);

$k := 0$; % lặp

$p_0 := 1 \in \mathbb{R}^n$; % vector khởi đầu

repeat

$k := k + 1$;

$p_k := A^T p_{k-1}$; % đánh giá vector riêng

$i := \operatorname{argmax}_j \{p_k[j]\}$; % chỉ mục giá trị lớn nhất

$\lambda := p_k[i] / p_{k-1}[i]$; % đánh giá giá trị riêng

$p_k := (1 / p_k[i]) * p_k$; % vector nhân lên

until $\|p_k - p_{k-1}\| \leq \epsilon$;

$p := (1 / \|p_k\|) * p_k$; % vector riêng chuẩn hóa

return (p, λ).

Xét đồ thị vừa cho. Bắt đầu từ vector uy tín khởi đầu $p_0 = (1, 1, 1, 1, 1)^T$, tính được vài bước lặp theo phương pháp lũy thừa, để tính vector riêng trội của A^T . Trong mỗi lần lặp, được $p_k = A^T p_{k-1}$.

Bảng 4.11. Phương pháp lũy thừa qua các lần nhân

p_0	p_1	p_2	p_3
$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 1 \\ 1 \\ 0.5 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1.5 \\ 1.5 \\ 0.5 \\ 1.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.67 \\ 1 \\ 1 \\ 0.33 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1.33 \\ 1.33 \\ 0.67 \\ 1.33 \end{pmatrix} \rightarrow \begin{pmatrix} 0.75 \\ 1 \\ 1 \\ 0.5 \\ 1 \end{pmatrix}$
λ	2	1.5	1.33

¹ Dominant eigenvector: vector riêng trội, đặc trưng nổi bật.

p_4	p_5	p_6	p_7
$\begin{pmatrix} 1 \\ 1.5 \\ 1.5 \\ 0.75 \\ 1.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.65 \\ 1 \\ 1 \\ 0.5 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1.5 \\ 1.5 \\ 0.67 \\ 1.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.67 \\ 1 \\ 1 \\ 0.44 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1.44 \\ 1.44 \\ 0.67 \\ 1.44 \end{pmatrix} \rightarrow \begin{pmatrix} 0.69 \\ 1 \\ 1 \\ 0.46 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1.46 \\ 1.46 \\ 0.69 \\ 1.46 \end{pmatrix} \rightarrow \begin{pmatrix} 0.68 \\ 1 \\ 1 \\ 0.47 \\ 1 \end{pmatrix}$
1.5	1.5	1.444	1.462

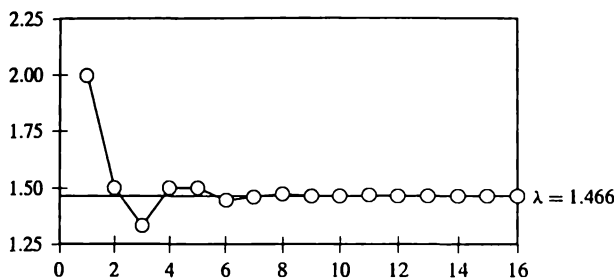
Chẳng hạn tính:

$$P_1 = A^T p_0 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \end{pmatrix}.$$

Trước khi sang lần lặp tiếp, lấy chia p_1 , mỗi mục, cho giá trị cực đại trong vector, tức 2, được:

$$p_1 = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \\ 1 \\ 0.5 \\ 1 \end{pmatrix}.$$

Khi k lớn, thực hiện $p_k = A^T p_{k-1} \approx \lambda p_{k-1}$, ngầm định rằng tỉ lệ của phần tử lớn nhất của p_k với p_{k-1} sẽ gần với λ . Bảng trên thể hiện tỉ lệ này. Hình sau cho thấy tỉ lệ này hội tụ về $\lambda = 1.466$, sau 10 lần lặp.



Hình 4.19. Hội tụ của tỉ lệ λ về giá trị riêng trội

Vector riêng trội hội tụ về p_k . Sau khi chuẩn hóa nó về vector đơn vị, vector riêng trội là p .

$$p_k = \begin{pmatrix} 1 \\ 1.466 \\ 1.466 \\ 0.682 \\ 1.466 \end{pmatrix} \quad p = \begin{pmatrix} 0.356 \\ 0.521 \\ 0.521 \\ 0.243 \\ 0.521 \end{pmatrix}$$

Về uy tín, v_2 , v_3 và v_5 có giá trị lớn nhất, do chúng có bậc trong là 2 và được các nút trò đến với cùng giá trị uy tín. Mặt khác, dù v_1 và v_4 có cùng bậc trong, v_1 được xếp hạng cao hơn, do v_3 đã mang uy tín của nó đến v_1 , còn v_4 chỉ có uy tín từ v_1 .

Xếp hạng: *Xếp hạng*¹ là phương pháp tính uy tín hay hướng tâm của các nút trong ngữ cảnh tìm kiếm Web. Đồ thị Web gồm các trang, tức các nút, nối nhau theo siêu liên kết, tức theo các cạnh. Phương pháp này sử dụng *lướt ngẫu nhiên*, giả sử người lướt Web chọn ngẫu nhiên một trong các đường liên kết ra từ trang hiện tại, hay nhảy ngẫu nhiên với xác suất rất nhỏ đến trang bất kì khác trên đồ thị Web.

Xếp hạng của một trang Web được xác định theo xác suất để người lướt Web chọn trang này. Giống như uy tín, xếp hạng của nút v phụ thuộc đệ quy vào xếp hạng của các nút khác trò đến nó.

¹ PageRank: xếp hạng trang Web.

Uy tín chuẩn hóa: giả sử mỗi nút u có cạnh ra ít nhất là 1. Người ta đề cập cách quản lí các nút không có cạnh ra sau. Lấy $od(u) = \sum_v A(u, v)$ kí hiệu bậc ra của nút u . Do việc lướt ngẫu nhiên có thể chọn liên kết ra bất kì, nếu có liên kết từ u đến v , thì xác suất thăm v từ u là $1/od(u)$.

Bắt đầu từ xác suất khởi đầu, hay xếp hạng $p_0(u)$ đối với mỗi nút, mà $\sum_u p_0(u) = 1$, có thể tính và cập nhật vector xếp hạng đối với v như sau: $p(v) = \sum_u \frac{A(u, v)}{od(u)} p(u) = \sum_u N(u, v) * p(u) = \sum_u N^T(v, u) * p(u)$, trong đó N là ma trận kề chuẩn hóa của đồ thị này. $N(u, v) = 1/od(u)$ nếu $(u, v) \in E$, 0 nếu ngược lại. Duyệt các nút, người ta có vector xếp hạng $p' = N^T p$. Vector xếp hạng là vector uy tín đã chuẩn hóa.

Nhảy ngẫu nhiên: Theo tiếp cận nhảy ngẫu nhiên, có xác suất rất nhỏ để nhảy từ nút này đến nút khác trong đồ thị, cho dù không có liên kết giữa chúng. Người ta có thể nghĩ về đồ thị Web như đồ thị có hướng đầy đủ, ảo, với ma trận kề $A_r = 1_{n \times n}$; trong đó $1_{n \times n}$ là ma trận có các phần tử là 1. Đối với ma trận lướt ngẫu nhiên, bậc ra của mỗi nút là $od(u) = n$, và xác suất nhảy từ nút u đến nút v bất kì là $1/od(u) = 1/n$. Vậy nếu chỉ cho phép nhảy ngẫu nhiên từ một nút đến nút khác, xếp hạng có thể tính tương tự như trước:

$$p(v) = \sum_u \frac{A_r(u, v)}{od(u)} p(u) = \sum_u N_r(u, v) * p(u) = \sum_u N_r^T(v, u) * p(u);$$

trong đó N_r là ma trận kề chuẩn hóa của đồ thị Web liên thông. N_r là ma trận gồm toàn phần tử $1/n$; $N_r = A_r/n = 1_{n \times n}/n$. Duyệt tất cả các nút, vector xếp hạng nhảy ngẫu nhiên được thể hiện theo $p' = N_r^T p$.

Xếp hạng: Xếp hạng đầy đủ được tính, khi giả thiết có xác suất nhỏ nào đó α cho phép nhảy Web từ nút hiện thời u đến nút v ngẫu nhiên khác, với xác suất $1 - \alpha$ cho phép đi theo đường liên kết đã có từ u đến v . Nói cách khác, người ta kết hợp vector uy tín chuẩn hóa và vector nhảy ngẫu nhiên, được vector xếp hạng, tức $p' = (1 - \alpha)N^T p + \alpha N_r^T p = ((1 - \alpha)N^T + \alpha N_r^T)p = M^T p$; trong đó

$M = (1 - \alpha)N + \alpha N_r$ là ma trận kề chuẩn hóa được kết hợp, vector xếp hạng gán p_0 và cập nhật sau mỗi lần lặp. Có vấn đề nhỏ, khi nút u không có cạnh ra nào, tức $od(u) = 0$. Vậy khi $od(u) = 0$ thì dòng ứng với u trong M , kí hiệu là M_u , lấy $\alpha = 1$, tức $M_u = M_u$ nếu $od(u) > 0$; $(1/n)1_n T$, nếu ngược lại, với 1_n là vector n chiều với các phần tử là 1. Thuật toán lặp lũy thừa cũng dùng để tính vector riêng trội của M^T .

Xét đồ thị vừa giới thiệu trên. Ma trận kề chuẩn hóa là N .

$$N = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0 & 0.33 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad N_r = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

Có $n = 5$ nút trong đồ thị, ma trận kề ngẫu nhiên là N_r . Giả sử $\alpha = 0.1$, ma trận kề chuẩn hóa là M . Tính vector riêng trội và giá trị riêng của M^T thu được $\lambda = 1$ và có p . Nút v_2 có giá trị xếp hạng cao nhất.

$$M = 0.9 * N + 0.1 * N_r = \begin{pmatrix} 0.02 & 0.02 & 0.02 & 0.92 & 0.02 \\ 0.02 & 0.02 & 0.47 & 0.02 & 0.47 \\ 0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.32 & 0.32 & 0.02 & 0.32 \\ 0.02 & 0.92 & 0.02 & 0.02 & 0.02 \end{pmatrix}, \quad p = \begin{pmatrix} 0.419 \\ 0.546 \\ 0.417 \\ 0.422 \\ 0.417 \end{pmatrix}$$

Điểm được phép và điểm trung tâm: xếp hạng của một nút độc lập với bất kì câu hỏi của người dùng, vì nó là giá trị chung đối với trang Web. Tuy nhiên, đối với câu hỏi khoa học, trang Web có xếp hạng chung cao có thể không thích hợp. Người ta muốn có khái niệm tập trung vào câu hỏi, liên quan đến xếp hạng hay uy tín của trang Web. Phương pháp tìm kiếm chủ đề theo siêu liên kết HITS¹ được thiết kế nhằm mục đích này.

¹ HITS Hyperlink Induced Topic Search: tìm kiếm chủ đề theo siêu liên kết.

Thực ra, người ta tính hai giá trị để kiểm chứng tầm quan trọng của trang Web (i) *điểm được phép*¹ của trang tương tự như xếp hạng, hay uy tín, phụ thuộc vào nhiều trang tốt trò đến nó ra sao; (ii) *điểm tập trung*² dựa vào nhiều ra sao các trang tốt mà nó trò đến. Nói khác đi, một trang được phép cao có nhiều trang tập trung trò đến, và trang có điểm tập trung cao sẽ trò đến nhiều trang được phép cao.

Với câu hỏi người dùng đã biết, phương pháp HITS dùng bộ tìm kiếm chuẩn để tìm tập các trang thích hợp. Rồi nó nói rộng tập này để chứa bất kì trang nào mà trò đến trang trong tập, hay bất kì trang nào được trò đến từ trang trong tập này. Bất kì trang nào xuất phát từ cùng trang chủ sẽ bị loại. HITS được áp dụng chỉ trên đồ thị câu hỏi mở rộng.

Kí hiệu $a(u)$ là điểm được phép và $h(u)$ là điểm tập trung của nút u . Điểm được phép phụ thuộc vào điểm tập trung và ngược lại, theo công thức sau: $a(v) = \sum_u A^T(v, u) * h(u)$; $h(u) = \sum_v A^T(v, u) * a(v)$. Theo kí pháp ma trận, có $a' = A^T h$, $h' = A a$. Thực ra, có thể viết lại theo đệ quy $a_k = A^T h_{k-1} = A^T (A a_{k-1}) = (A^T A) a_{k-1}$; và $h_k = A a_k = A (A^T h_{k-1}) = (A A^T) h_{k-1}$.

Khi $n \rightarrow \infty$, điểm được phép hội tụ về vectơ riêng trội của $A^T A$, và điểm tập trung hội tụ về vectơ riêng của $A A^T$. Phương pháp lặp lũy thừa có thể được sử dụng để tính vectơ riêng trong cả hai trường hợp. Bắt đầu với vectơ $a = 1_n$, tức vectơ toàn 1, tính được vectơ $h = A a$. Để tránh tràn ô nhớ, người ta chia vectơ cho phần tử cực đại. tiếp theo, tính $a = A^T h$, và cũng chia cho phần tử cực đại. Quá trình lặp cho đến khi a và h hội tụ.

Đối với đồ thị đã nêu trước đây, có thể tính lặp điểm được phép và điểm tập trung, với khởi đầu $a = (1, 1, 1, 1, 1)^T$. Trong lần lặp đầu, được h . Sau đó làm lại thang đo, bằng cách chia cho giá trị lớn nhất, được h' .

¹ Authority score: điểm được phép.

² Hub score: điểm tập trung.

$$h = Aa = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 1 \end{pmatrix}, \quad h' = \begin{pmatrix} 0.33 \\ 0.67 \\ 0.33 \\ 1 \\ 0.33 \end{pmatrix}$$

Tiếp theo, cập nhật a , rồi chia cho giá trị lớn nhất, được a' . Quá trình tiếp tục cho đến khi a , h hội tụ về vector riêng trội của $A^T A$ và AA^T , được a và h .

$$a = A^T h' = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.33 \\ 0.67 \\ 0.33 \\ 1 \\ 0.33 \end{pmatrix} = \begin{pmatrix} 0.33 \\ 1.33 \\ 1.67 \\ 0.33 \\ 1.67 \end{pmatrix}$$

$$a' = \begin{pmatrix} 0.2 \\ 0.8 \\ 1 \\ 0.2 \\ 1 \end{pmatrix} \quad a = \begin{pmatrix} 0 \\ 0.46 \\ 0.63 \\ 0 \\ 0.63 \end{pmatrix} \quad h = \begin{pmatrix} 0 \\ 0.58 \\ 0 \\ 0.79 \\ 0.21 \end{pmatrix}$$

Từ các điểm số này, kết luận rằng v_4 có điểm tập trung lớn nhất, do nó trò đến ba nút, v_2 , v_3 và v_5 với điểm được phép tốt. Nói cách khác, cả v_3 và v_5 có điểm được phép cao, do hai nút v_4 và v_2 với điểm tập trung cao trò đến chúng.

4.3.4. Các mô hình đồ thị

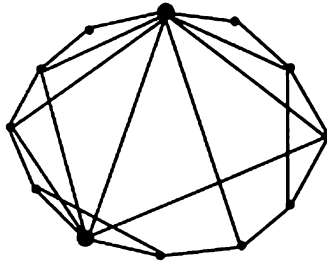
4.3.4.1. Giới thiệu về mô hình đồ thị

Nhiều mạng trong thực tế thể hiện các đặc tính chung, cho dù dữ liệu bên dưới thuộc các lĩnh vực đa dạng, chẳng hạn như mạng xã hội, mạng sinh học, mạng truyền thông... Vấn đề chung đối với các mạng là hiểu quá trình bên dưới chúng. Các mạng thể giới thực thường *lớn* và *thưa*, tức cấp hay số các nút, n , là rất lớn, và

kích thước của đồ thị, hay số cạnh, $m = O(n)$. Các mô hình đồ thị trong phần này đề cập đối tượng lớn và thưa.

Tính chất về thế giới nhỏ:

Định nghĩa: Mạng thế giới nhỏ¹ là đồ thị toán học, trong đó hầu hết các nút không phải là láng giềng của nhau, nhưng các láng giềng của bất kỳ nút xác định có thể là láng giềng của nhau và hầu hết các nút có thể được đến từ mọi nút khác bởi số nhỏ bước nhảy hoặc một số bước.



Hình 4.20. Mạng thế giới nhỏ; nút trung tâm lớn hơn các nút khác

Các đồ thị thế giới thực thể hiện tính chất *thế giới nhỏ*, tức có đường ngắn giữa bất kỳ cặp nút nào.

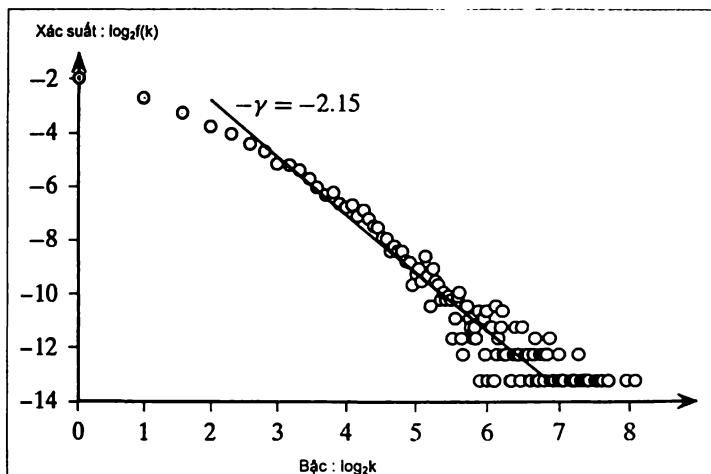
Định nghĩa: Đồ thị G thể hiện hành vi thế giới nhỏ nếu độ dài trung bình μ tỉ lệ theo logarit với số các nút n trong G , tức $\mu \propto \log n$.

Đồ thị có tính chất thế giới cực nhỏ nếu độ dài đường đi trung bình nhỏ hơn nhiều so với $\log n$, tức $\mu \ll \log n$.

Tính chất tự do về tỉ lệ: trong nhiều đồ thị thực, người ta thấy phân phối bậc thực nghiệm $f(k)$ thể hiện hành vi tự do về bậc, có quan hệ lũy thừa với k , tức xác suất nút có bậc k thỏa mãn điều kiện $f(k) \propto k^{-\gamma}$. Về trực quan, luật lũy thừa cho biết phần lớn các nút có bậc rất nhỏ, trong khi có một ít các nút trung tâm với bậc cao, tức chúng liên kết, tương tác, với nhiều nút. Mỗi quan hệ theo luật lũy thừa dẫn đến hành vi tự do về bậc, hay bất biến, do lấy tỉ

¹ Small - world network: mạng thế giới nhỏ.

lệ lại theo hằng số c sẽ không thay đổi tỉ lệ giữa chúng. Viết lại $f(k) \propto k - \gamma$ với hằng số tỉ lệ α , được $f(k) \propto \alpha k - \gamma$. Vậy $f(ck) = \alpha (\alpha k) - \gamma = (\alpha c - \gamma)k - \gamma \propto k - \gamma$. Lấy logarit $f(k) \propto \alpha k - \gamma$, được $\log f(k) = \log (\alpha k - \gamma)$, hay $\log f(k) = -\gamma \log k + \log \alpha$, cho quan hệ logarit giữa k và $f(k)$, như trong hình vẽ sau.

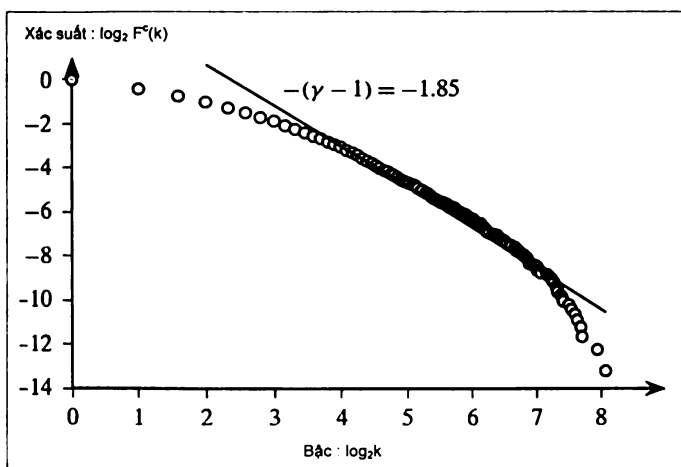


Hình 4.21. Phân phối bậc

Trong thực tế, một trong những vấn đề về đánh giá phân phối bậc đối với đồ thị là mức cao của ồn đối với các bậc cao, khi số đếm tần số là nhỏ nhất. Một tiếp cận giải quyết vấn đề này là sử dụng phân phối bậc tích lũy $F(k)$, sẽ làm mịn các ồn. Thực tế người ta sử dụng $F^c(k) = 1 - F(k)$, cho xác suất để nút được ngẫu nhiên có bậc lớn hơn k . Nếu $f(k) \propto k - \gamma$ và giả sử $\gamma > 1$, thì $F^c(k) = 1 - F(k)$

$$= 1 - \sum_{x=0}^k f(x) = \sum_{x=k}^{\infty} f(x) = \sum_{x=k}^{\infty} x^{-\gamma} \approx \int_k^{\infty} x^{-\gamma} dx = \frac{x^{-\gamma+1}}{-\gamma+1} \Big|_k^{\infty} = \frac{1}{\gamma-1} k^{-(\gamma-1)} \propto k^{-(\gamma-1)}$$

Nói khác đi, đường log - log của $F^c(k)$ so với k sẽ theo luật lũy thừa với độ dốc $-(\gamma - 1)$, khác với $-\gamma$. Với hiệu ứng làm mịn, người ta có đánh giá luật lũy thừa tốt hơn.



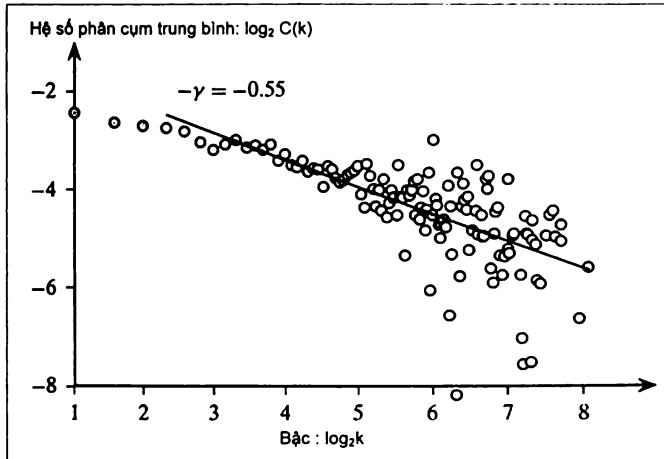
Hình 4.22. Phân phối bậc tích lũy

Hiệu ứng phân cụm: đồ thị thế giới thực thường thể hiện *hiệu ứng phân cụm*, tức hai nút có vẻ liên thông nếu chúng có chung láng giềng. Hiệu ứng phân cụm thể hiện qua hệ số phân cụm cao đối với đồ thị G . Kí hiệu $C(k)$ là hệ số phân cụm trung bình đối với tất cả các nút với bậc k , thì hiệu ứng phân cụm cũng tự thể hiện mối quan hệ luật lũy thừa giữa $C(k)$ và k : $C(k) \propto k^{-\gamma}$. Về trực quan, hành vi luật lũy thừa cho biết phân cụm phân cấp của các nút này. Tức các nút hiếm khi liên thông, tức có bậc nhỏ hơn, là một phần của các miền đã phân cụm cao, tức có hệ số phân cụm cao hơn. Ngoài ra, chỉ một ít các nút trung tâm, với bậc cao, nổi với các miền đã phân cụm này; các nút trung tâm có hệ số phân cụm nhỏ hơn.

Hình vẽ trên cho thấy phân phối bậc liên quan đến tương tác *protein*¹ của người. Nút ứng với protein, cung ứng với tương tác theo thực nghiệm của hai protein kề nhau. Đồ thị có $n = 9521$ nút và $m = 37.060$ cung. Mối quan hệ tuyến tính giữa $\log k$ và $\log f(k)$ là rõ ràng, dù các giá trị bậc rất lớn hay rất nhỏ không phù hợp với xu thế. Đường khít nhất sau khi bỏ các bậc cực trị có giá trị $\gamma = 2.15$.

¹ Protein: còn gọi là chất đạm, là những đại phân tử được cấu tạo theo nguyên tắc đa phân mà các đơn phân là axit amin.

Đường logk so với $\log F^c(k)$ nổi rõ là tuyến tính. Độ dốc là $-(\gamma - 1) = 1.85$, tức $\gamma = 2.85$. Người ta có thể kết luận rằng đồ thị này thể hiện hành vi tự do về tỉ lệ, trừ các bậc cực trị, với γ giữa 2 và 3, như nhiều đồ thị thế giới thực.



Hình 4.23. Phân phối hệ số phân cụm trung bình

Đường kính của đồ thị là $d(G) = 14$, rất gần với $\log_2 n = \log_2(9621) = 13.22$. Vậy nên mạng này là mạng rất nhỏ. Hình 4.23 vẽ hệ số phân cụm trung bình như hàm của bậc. Đường log - log có xu thế tuyến tính yếu; đường phù hợp nhất có độ dốc $-\gamma = -0.55$. Người ta có thể kết luận rằng đồ thị thể hiện hành vi phân cụm phân cấp yếu.

4.3.4.2. Mô hình đồ thị Erdos - Rényi

Định nghĩa: Mô hình ER^1 tạo đồ thị ngẫu nhiên, mà bất kì đồ thị có số các nút và cạnh xác định có xác suất được chọn như nhau.

Mô hình ER có hai tham số (i) số các nút n ; (ii) số cạnh m . Ký hiệu M là số cực đại các cạnh từ n nút, tức $M = \binom{n}{2} = n(n-1)/2$.

¹ ER: mô hình do Paul Erdős và Alfréd Rényi, đưa ra năm 1959.

Mô hình này xác định bộ các đồ thị $CG(n, m)$, sao cho mỗi đồ thị $G \in CG$ đều có xác suất được chọn là $P(G) = \frac{1}{\binom{M}{m}} = \left(\frac{M}{m}\right)^{-1}$; trong đó $\binom{M}{m}$ là số các đồ thị có thể có m cạnh, n nút, ứng với cách chọn m từ M cạnh.

Kí hiệu $V = \{v_1, \dots, v_n\}$ là tập n nút. Phương pháp ER chọn ngẫu nhiên đồ thị $G = (V, E) \in CG$ nhờ quá trình sinh. Tại mỗi bước, nó chọn ngẫu nhiên hai đỉnh phân biệt $v_i, v_j \in V$, và thêm cạnh (v_i, v_j) vào E , đảm bảo cạnh này chưa có trong G . Quá trình lặp đến khi chính xác m cạnh được bổ sung vào đồ thị này.

Kí hiệu biến ngẫu nhiên X là bậc của nút, để $G \in CG$. Kí hiệu p là xác suất của một cạnh trong G , được tính theo $p = m/M = m/\binom{n}{2} = 2m/(n(n-1))$.

Bậc trung bình: Đối với nút đã biết bất kì trong G , bậc của nó có thể lớn nhất, là $n-1$, do không cho phép vòng lặp. Do p là xác suất để một cạnh từ nút bất kì, nên biến ngẫu nhiên X , ứng với bậc của nút, tuân theo phân phối nhị thức với xác suất thành công là p , $f(k) = P(X = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$. Bậc trung bình μ_d như giá trị kì vọng của X : $\mu_d = E[X] = (n-1)p$.

Người ta có thể tính phương sai của bậc theo các nút, nhờ tính phương sai của X : $\sigma_d^2 = \text{var}(X) = (n-1)p(1-p)$.

Phân phối bậc: Để có phân phối bậc được đồ thị ngẫu nhiên thưa và lớn, cần rút ra biểu thức đối với $f(k) = P(X = k)$ khi $n \rightarrow \infty$. Giả sử $m = O(n)$, có thể viết $p = \frac{m}{n(n-1)/2} = \frac{O(n)}{n(n-1)/2} = \frac{1}{O(n)} \rightarrow 0$. Nói khác đi, người ta quan tâm đến hành vi bất đối xứng của đồ thị, khi $n \rightarrow \infty$, và $p \rightarrow 0$.

Dưới hai xu thế này, giá trị kì vọng và phương sai của X được viết lại: $E[X] = (n - 1)p \approx np$ khi $n \rightarrow \infty$; $\text{var}(X) = (n - 1)p(n - 1) \approx np$ khi $p \rightarrow 0$. Nói cách khác, đối với đồ thị thưa và lớn, kì vọng và phương sai của X là như nhau, $E[X] = \text{var}(X) = np$, và phân phối nhị thức có thể xấp xỉ nhờ phân phối Poisson với tham số λ , với $f(k) = \lambda^k \cdot e^{-\lambda}/k!$; trong đó $\lambda = np$ thể hiện cả giá trị kì vọng lẫn phương sai của phân phối này. Sử dụng xấp xỉ Stirling: $k! \approx k^k e^{-k} \sqrt{2\pi k}$, được $f(k) = \frac{\lambda^k e^{-k}}{k!} = \frac{\lambda^k e^{-k}}{k^k e^{-k} \sqrt{2\pi k}} = \frac{e^{-k}}{\sqrt{2\pi k}} * \frac{(\lambda e)^k}{\sqrt{k} * k^k}$. Nói

khác đi, người ta có $f(k) \propto \alpha^k k^{-\frac{1}{2}} k^{-k}$, với $\alpha = \lambda e = npe$. Kết luận rằng đồ thị thưa và lớn theo phân phối bậc Poisson, chứ không theo mối quan hệ luật lũy thừa. Vậy trong trường hợp này, mô hình đồ thị ER không đủ để miêu tả đồ thị tự do về tỉ lệ của thế giới thực.

Hệ số phân cụm: Xét nút v_i trong đồ thị G với bậc k . Hệ số phân cụm của v_i là $C(v_i) = 2m_i / (k(k - 1))$, trong đó $k = n_i$ là kí hiệu số nút, m_i kí hiệu số cạnh trong đồ thị con sinh ra từ láng giềng của v_i . Tuy nhiên, do p là xác suất chọn một cạnh, số kì vọng để m_i trong các láng giềng của v_i là $m_i = p_k (k - 1) / 2$. Vậy $C(v_i) = 2m_i / (k(k - 1)) = p$. Nói khác đi, hệ số phân cụm kì vọng qua tất cả các nút của tất cả các bậc là thống nhất, nên hệ số phân cụm toàn thể cũng là thống nhất $C(G) = \frac{1}{n} \sum_i C(v_i) = p$.

Ngoài ra, đối với các đồ thị thưa, người ta có $p \rightarrow 0$; điều này ngầm định $C(G) = C(v_i) \rightarrow 0$. Do đó đồ thị ngẫu nhiên lớn không có hiệu ứng phân cụm, ngược với nhiều mạng thế giới thực.

4.3.4.3. Đường kính

Bậc kì vọng của một nút là $\mu_d = \lambda$, có nghĩa trong một lần nhảy từ nút đã biết, có thể đến λ nút khác. Do mỗi láng giềng của một nút khởi đầu cũng có bậc trung bình λ , người ta có thể xấp xỉ số các nút mà hai lần nhảy là λ^2 . Nói chung, với mức đánh giá thô, tức bỏ qua các láng giềng chung, người ta có thể đánh giá số các nút tại khoảng cách k lần nhảy khỏi nút ban đầu v_i như λ^k . Do có

tất cả n đỉnh phân biệt trong đồ thị, nên $\sum_{k=1}^t \lambda^k = n$; trong đó t là số cực đại của bước nhảy từ v_i . Người ta có $\sum_{k=1}^t \lambda^k = \frac{\lambda^{t+1} - 1}{\lambda - 1} \approx \lambda^t$, nên $\lambda^t \approx n$, hay $t \log \lambda \approx \log n$, ngầm định $t \approx \log n / \log \lambda \propto \log n$. Do độ dài của đường từ một nút đến nút xa nhất bị t chặn, đường kính của đồ thị cũng bị chặn, tức $d(G) \propto \log n$, với giả thiết bậc kì vọng λ là cố định. Người ta có thể kết luận được đồ thị ngẫu nhiên thỏa mãn ít nhất một tính chất của đồ thị thế giới thực, gọi là chúng thể hiện hành vi thế giới thực.

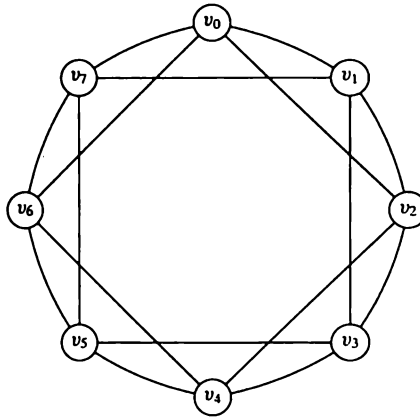
4.3.4.4. Mô hình đồ thị thế giới nhỏ Watt - Strogatz

Mô hình đồ thị ngẫu nhiên không phù hợp để thể hiện hệ số phân cụm cao, nhưng đó là thế giới nhỏ. Mô hình WS¹ mô hình hóa tường minh việc phân cụm cục bộ cao, bắt đầu với mạng chính quy mà mỗi nút là nối với k láng giềng về bên phải, bên trái, với giả thiết n đỉnh khởi đầu được sắp xếp theo đường trục xương sống vòng quanh lớn. Mạng như vậy sẽ có hệ số phân cụm cao, nhưng sẽ không là thế giới nhỏ. Khi thêm lượng nhỏ ngẫu nhiên trong mạng chính quy nhờ liên kết ngẫu nhiên vài cạnh, hay khi thêm phần nhỏ các cạnh ngẫu nhiên, dẫn đến hiện tượng thế giới nhỏ.

Mô hình WS bắt đầu với n nút đã xếp vòng quanh, với mỗi nút liên thông với láng giềng trái, phải bên cạnh. Các cạnh trong vùng khởi đầu được gọi là các cạnh *xương sống*; mỗi nút có các cạnh đến $k - 1$ láng giềng thêm vào, về bên trái và bên phải. Vậy, mô hình WS bắt đầu với đồ thị *chính quy* bậc $2k$, trong đó mỗi nút nối với k láng giềng về bên trái, như trong hình sau.

Hệ số phân cụm và đường kính của đồ thị chính quy: xét đồ thị con G_v sinh ra từ $2k$ láng giềng của v . Hệ số phân cụm của v là $C(v) = m_v / M_v$; trong đó m_v là số cạnh hiện thời, M_v là số cạnh nhiều nhất có thể trong số các láng giềng của v .

¹ WS do Duncan J. Watts và Steven Strogatz đề xuất, năm 1998.



Hình 4.24. Đồ thị chính quy Watts - Strogatz, $n = 8, k = 2$

Để tính m_v , xét nút v_i cách xa i bước nhảy từ v về bên phải, chỉ trên các cạnh xương sống. Nút v_i có cạnh đến $k - i$ láng giềng ngay bên phải, và $k - 1$ láng giềng bên trái, với tất cả k láng giềng trái, trừ v . Do đối xứng với v , nút v_i cách xa i bước nhảy từ v có cùng số cạnh. Vậy, bậc của bất kì nút nào trong G_v cách i bước nhảy từ v là $d_i = (k - i) + (k - 1) = 2k - i - 1$. Do mỗi cạnh tham gia vào bậc của hai nút kề, tổng các bậc của tất cả các láng giềng của v : $2m_v = 2 \sum_{i=1}^k 2k - i - 1$ nên $m_v = 2k^2 - k(k + 1)/2 - k = 3k(k - 1)/2$.

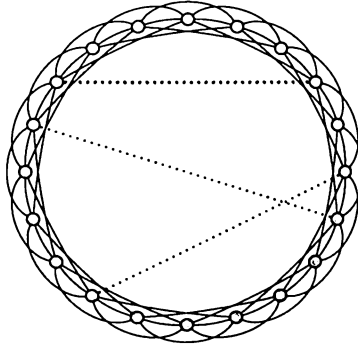
Số các cạnh có thể trong số $2k$ láng giềng của v là $M_v = \binom{2k}{2} = 2k$

$(2k - 1)/2 = k(2k - 1)$. Do vậy $C(v) = m_v / M_v = (3k - 3) / (4k - 2)$. Khi k tăng, hệ số phân cụm dần về $3/4$, do $C(G) = C(v) \rightarrow 3/4$, khi $k \rightarrow \infty$.

Vậy đồ thị chính quy WS có hệ số phân cụm cao. Tuy nhiên, nó không thỏa mãn tính chất thế giới nhỏ. Theo đường xương sống, nút xa nhất từ v có khoảng cách nhiều nhất $n/2$ bước nhảy. Ngoài ra, do mỗi nút nối với k láng giềng về mỗi bên, nó đến nút xa nhất nhiều nhất $(n/2)/k$ bước nhảy. Chính xác thì đường kính của đồ thị chính quy WS là $d(G) = \lceil n / (2k) \rceil$ nếu n chẵn; $\lceil (n - 1) / (2k) \rceil$ nếu ngược lại. Đồ thị chính quy có đường kính tỉ lệ tuyến tính với số nút, nên không là thế giới nhỏ.

Nhiều loạn ngẫu nhiên của đồ thị chính quy: Xét hai tiếp cận sau.

1. *Nối lại cạnh.* Bắt đầu với đồ thị chính quy có bậc $2k$, mô hình WS làm nhiều cấu trúc chính quy bằng cách thêm vài ngẫu nhiên vào mạng. Một tiếp cận là nối lại ngẫu nhiên các cạnh với xác suất r . Tức đối với mỗi cạnh (u, v) trong đồ thị, với xác suất r , thay v bằng nút chọn ngẫu nhiên khác, tránh xoay vòng và lặp lại cạnh. Do đồ thị chính quy WS có tất cả $m = kn$ cạnh, sau khi nối lại, rm cạnh là ngẫu nhiên, và $(1 - r)m$ là bình thường;
2. *Đường tắt của cạnh.* Tiếp cận khác là thay vì nối lại cạnh, người ta thêm *cạnh nối tắt* giữa cặp nút ngẫu nhiên. Số tổng cộng các cạnh nối tắt được thêm vào mạng là $mr = knr$, mà r là xác suất, đối với cạnh, bổ sung cạnh nối tắt. Khi đó số cạnh tổng cộng trong đồ thị là $m + mr = (1 + r)m = (1 + r)kn$. Do $r \in [0; 1]$, số các cạnh vào khoảng $[kn, 2kn]$.



Hình 4.25. Đồ thị Watts - Strogatz, $n = 20, k = 3$: cạnh nối tắt chấm chấm

Trong mỗi tiếp cận, nếu xác suất nối lại hay nối tắt cạnh $r = 0$, thì người ta trở về đồ thị chính quy, với hệ số phân cụm cao, nhưng không có tính chất thế giới nhỏ. Mặt khác, nếu $r = 1$, cấu trúc chính quy bị phá, và đồ thị này trở thành đồ thị ngẫu nhiên, ít có hiệu ứng phân cụm, nhưng có tính chất thế giới nhỏ. Như vậy, tác động ngẫu nhiên nhỏ cũng đã thay đổi mạng chính quy. Như trong hình minh họa, chỉ với giá trị r thấp, mô hình WS giữ

lại hầu hết cấu trúc phân cụm bình thường, nhưng trở nên mạng thế giới nhỏ.

Thuộc tính của đồ thị Watts – Strogatz: thuộc tính phân phối bậc: xét tiếp cận nối tắt cạnh, tiện cho phân tích. Mỗi đỉnh có bậc ít nhất là $2k$. Ngoài ra, có đường tắt, theo phân phối nhị thức. Mỗi nút có $n' = n - 2k - 1$ cạnh nối tắt thêm vào, vậy lấy n' như số phép thử độc lập để bổ sung cạnh. Do một nút có bậc $2k$, với xác suất cạnh nối tắt là r , kì vọng thô $2kr$ đường nối tắt từ nút đó, nhưng nút này có thể nối nhiều nhất $n - 2k - 1$ nút khác. Vậy, xác suất thành công là $p = 2kr / (n - 2k - 1) = 2kr / n'$.

Kí hiệu X là biến ngẫu nhiên về số cạnh nối tắt đối với mỗi nút. Xác suất của nút với j cạnh nối tắt là $f(j) = P(X = j) = \binom{n'}{j} p^j (1-p)^{n'-j}$, với $E[X] - n'p = 2kr$. Vậy, bậc kì vọng của mỗi nút trong mạng này là $2k + E[X] = 2k + 2kr = 2k(1+r)$. Rõ ràng là phân phối bậc của đồ thị WS không liên quan đến luật lũy thừa; vậy nên các mạng này không tự do về tỉ lệ.

Thuộc tính hệ số phân cụm: sau khi bổ sung các cạnh nối tắt, mỗi nút v có bậc kì vọng là $2k(1+r)$, tức về trung bình nó nối với $2kr$ láng giềng mới, ngoài $2k$ ban đầu. Số các cạnh có thể trong số láng giềng của v là $M_v = 2k(1+r)(2k(1+r)-1)/2 = (1+r)k(4kr+2k-1)$. Do đồ thị chính quy WS vẫn y nguyên, cho dù bổ sung các cạnh nối tắt, láng giềng của v giữ lại tất cả $3k(k-1)/2$ cạnh ban đầu. Ngoài ra, một vài cạnh nối tắt có thể liên kết cặp các nút trong số các láng giềng của v . Kí hiệu biến ngẫu nhiên Y là số cạnh nối tắt trong số $2k(1+r)$ láng giềng của v ; vậy Y theo phân phối nhị thức với xác suất thành công p . Do đó số kì vọng của các cạnh nối tắt là $E[Y] = pM_v$. Lấy m_v là biến ngẫu nhiên ứng với số cạnh hiện thời trong số các láng giềng của v , cả nối tắt lẫn bình thường. Số kì vọng của các cạnh trong láng giềng của v là

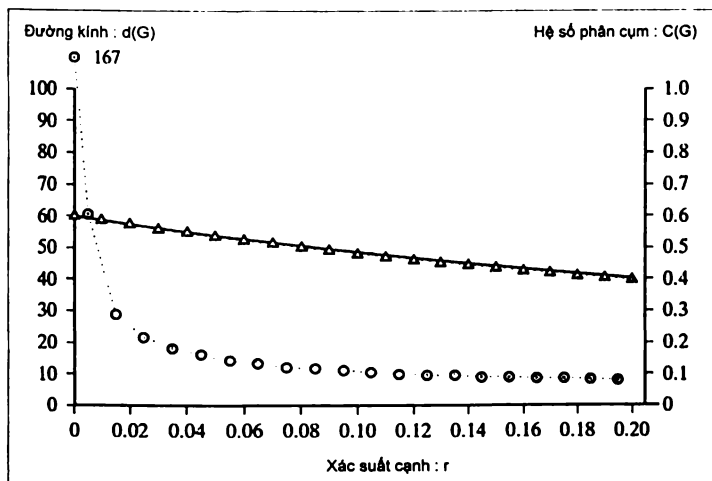
$$E[m_v] = E[3k(k-1)/2 + Y] = 3k(k-1)/2 + pM_v.$$

Do phân phối nhị thức là đặc biệt tập trung quanh trung bình, người ta có thể xấp xỉ hệ số phân cụm nhờ số cạnh kì vọng, theo

$C(v) \approx E[m_v]/M_v = (3k(k-1)/2 + pM_v)/M_v = 3k(k-1)/2M_v + p = 3(k-1)/((1+r)(4kr+2(2k-1))) = (3k-3)/(4k-2+2r(2k+4k-1))$.
 Khi $r \rightarrow \infty$, biểu thức này tương đương với công thức $C(v) = m_v/M_v$.
 Vậy đối với các giá trị r nhỏ, hệ số phân cụm vẫn cao.

Thuộc tính đường kính: việc phân tích biểu thức về đường kính của mô hình WS là không dễ dàng. Người ta nghiên cứu thực nghiệm về hành vi của đồ thị WS khi thêm ít ngẫu nhiên các cạnh nối tắt. Trong thí dụ sau, người ta biết rằng giá trị nhỏ về xác suất cạnh nối tắt r là đủ để rút ra đường kính từ $O(n)$ đến $O(\log n)$. Mô hình WS đưa đến đồ thị thế giới nhỏ và thể hiện hiệu quả phân cụm. Tuy nhiên, đồ thị WS không hiển thị phân phối bậc tự do về tỉ lệ.

Hình vẽ cho biết mô phỏng của mô hình WS, đối với đồ thị $n = 1000$ đỉnh và $k = 3$. Các giá trị đường kính hiện ra như các vòng tròn, trên trục X bên trái, còn giá trị phân cụm hiện ra như các tam giác với trục Y. Các giá trị này đã lấy trung bình trên 10 lần chạy mô hình WS. Đường liền ứng với hệ số phân cụm do phân tích công thức xấp xỉ $C(v)$, nhất trí với giá trị mô phỏng.



Hình 4.26. Mô hình Watts - Strogatz: đường kính (tròn) và hệ số phân cụm (tam giác)

Đồ thị chính quy khởi đầu có đường kính $d(G) = \lceil n/(2k) \rceil = \lceil 1000/6 \rceil = 167$ và hệ số phân cụm của nó là $C(G) = 3(k-1)/(2(2k-1)) = 6/10 = 0.6$. Người ta có thể quan sát thấy đường kính giảm nhanh, cho dù với xác suất bổ sung cạnh rất nhỏ. Với $r = 0.005$, đường kính là 61; với $r = 0.1$, đường kính quanh 11, như tỉ lệ $O(\log n)$, do $\log 1000 \approx 10$. Mặt khác, có thể quan sát thấy hệ số phân cụm còn cao. Với $r = 0.1$, hệ số phân cụm là 0.48. Vậy nghiên cứu thực nghiệm khẳng định rằng việc thêm cho dù số nhỏ các cạnh nối tắt ngẫu nhiên sẽ giảm đường kính của đồ thị chính quy WS, từ $O(n)$, tức thế giới lớn, xuống $O(\log n)$, tức thế giới nhỏ. Lúc này, đồ thị vẫn giữ thuộc tính phân cụm cục bộ.

4.3.4.5. Mô hình tự do về tỉ lệ Barabási - Albert

Mô hình BA¹ nắm được phân phối bậc tự do về tỉ lệ của đồ thị thế giới thực, thông qua quá trình sinh, bổ sung nút mới và cạnh mới tại mỗi bước. Ngoài ra, phát triển của cạnh dựa vào khái niệm *đỉnh kèm ưu tiên*, tức các cạnh từ một đỉnh mới có nhiều khả năng liên kết với các nút có bậc cao. Do nguyên nhân này, mô hình được gọi là *tiếp cận nước chảy chỗ trũng*². Mô hình AB bắt chước việc phát triển động đồ thị, nhờ bổ sung nút và cạnh mới tại mỗi bước thời gian $t = 1, 2, \dots$. Kí hiệu G_t là đồ thị tại thời điểm t , n_t là số các nút và m_t là số các cạnh của G_t .

Khởi động: Mô hình BA bắt đầu với $t = 0$, đồ thị khởi đầu là G_0 , với n_0 nút và m_0 cạnh. Mỗi nút trong G_0 có bậc ít nhất là 1; ngược lại nút này không bao giờ được chọn để *đỉnh kèm ưu tiên*. Người ta giả sử rằng mỗi nút có bậc khởi đầu là 2, liên kết sang láng giềng trái và phải trong đường vòng quanh. Vậy $m_0 = n_0$.

Phát triển và gắn ưu tiên: Mô hình BA suy ra đồ thị mới G_{t+1} từ G_t nhờ bổ sung chính xác một nút mới u , và bổ sung $q \leq n_0$ cạnh

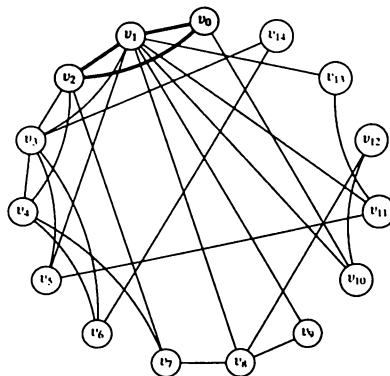
¹ Mô hình Barabási - Albert (BA) là một thuật toán để tạo các mạng tự do về tỉ lệ, ngẫu nhiên nhờ sử dụng cơ chế *đỉnh kèm ưu tiên*, do Albert - László Barabási và Réka Albert đề xuất năm 2002.

² Rich gets richer: giàu lại giàu thêm.

mới từ nút u đến q nút khác biệt $v_i \in G_t$; trong đó v_i được chọn với xác suất $\pi_t(v_i) = d_i / \sum_{v_i \in G_t} d_i$. Do chỉ một đỉnh mới được thêm vào ở mỗi bước, số nút trong G_t là $n_t = n_0 + t$. Ngoài ra, do thêm chính xác q cạnh mới, số cạnh trong G_t là máy tính $= m_0 + q_t$. Do tổng của các bậc là hai lần số cạnh trong đồ thị, nên $\sum_{v_i \in G_t} d(v_i) = 2m_t = 2(m_0 + q_t)$.

Viết lại biểu thức xác suất $\pi_t(v_i) = d_i / (2(m_0 + q_t))$. Do mạng phát triển và có đỉnh kèm ưu tiên, về trực quan người ta thấy xuất hiện bộ tập trung với bậc cao.

Hình vẽ cho thấy đồ thị được tạo theo mô hình BA, với tham số $n_0 = 3$, $q = 2$ và $t = 12$. Bắt đầu, với $t = 0$, đồ thị có $n_0 = 3$ đỉnh, gọi là $\{v_0, v_1, v_2\}$, nối theo $m_0 = 3$ cạnh, đường đậm. Tại bước $t = 1, \dots, 12$ đỉnh v_{t-2} được bổ sung và mạng phát triển và nối với $q = 2$ đỉnh được chọn với xác suất tỉ lệ theo bậc.



Hình 4.27. Đồ thị Barabási - Albert, $n_0 = 3$, $q = 2$, $t = 12$

Chẳng hạn, $t = 1$, đỉnh v_3 được thêm, với cạnh đến v_1 và v_2 , được chọn theo phân phối $\pi_0(v_i) = 1/3$, với $i = 0, 1, 2$. Tại $t = 2$, thêm v_4 . Nút v_2 và v_3 được chọn ưu tiên theo phân phối xác suất $\pi_1(v_0) = \pi_1(v_3) = 2/10 = 0.2$; $\pi_1(v_1) = \pi_1(v_2) = 3/10 = 0.3$. Đồ thị cuối cùng sau $t = 12$ cho thấy xuất hiện vài nút tập trung, như v_1 , bậc 9, và v_3 , bậc 6.

Phân phối bậc: Sau đây là hai tiếp cận khác nhau, để ước lượng phân phối bậc cho mô hình BA, là (i) tiếp cận rời rạc; (ii) tiếp cận liên tục.

Phân phối bậc theo tiếp cận rời rạc: Tiếp cận còn được gọi là phương pháp *phương trình chủ*¹. Kí hiệu biến ngẫu nhiên X_t là bậc của nút trong G_t , và $f_t(k)$ là hàm mật độ xác suất đối với X_t . Tức $f_t(k)$ là phân phối bậc cho đồ thị G_t tại bước thời gian t . Lấy đơn giản $f_t(k)$ là phần các nút với bậc k tại t . Kí hiệu n_t là số nút và máy tính số cạnh trong G_t . Ngoài ra, kí hiệu $n_t(k)$ là số nút với bậc k trong G_t . Vậy $f_t(k) = n_t(k)/n_t$. Do người ta quan tâm đến đồ thị thế giới thực lớn, khi $t \rightarrow \infty$, số các nút và cạnh trong G_t có thể xấp xỉ theo $n_t = n_0 + t \approx t$; $m_t = m_0 + q_t \approx q_t$. Theo công thức xác suất, tại bước $t + 1$, xác suất $\pi_t(k)$ để nút nào đó với bậc k trong G_t được chọn để gắn ưu tiên, được viết lại $\pi_t(k) = k \cdot n_t(k) / \sum_i i \cdot n_t(i)$. Chia tử số và mẫu số cho n_t , được $\pi_t(k) = k \cdot f_t(k) / \sum_i i \cdot f_t(i)$. Mẫu số là giá trị kì vọng của X_t , tức bậc có nghĩa trong G_t , do $E[X_t] = \mu_d(G_t) = \sum_i i \cdot f_t(i)$. Với bất kì đồ thị nào, bậc trung bình là $\mu_d(G_t) = \sum_i d_i / n_t = 2m_t / n_t \approx 2q_t / t = 2q$; trong đó theo nhận xét trên, $m_t = q_t$. Có thể viết lại xác suất gắn ưu tiên đối với nút bậc k là $\pi_t(k) = k \cdot f_t(k) / (2q)$.

Xét thay đổi về số nút với bậc k , khi nút mới n kết hợp với mạng phát triển tại bước thời gian $t + 1$. Thay đổi mạng về số với bậc k như số các nút với bậc k tại bước $t + 1$ trừ đi số các nút với bậc k tại bước t , nên $(n_t + 1) \cdot f_{t+1}(k) - n_t \cdot f_t(k)$. Dùng xấp xỉ $n_t \approx t$, thay đổi mạng về các nút bậc k là $(n_t + 1) \cdot f_{t+1}(k) - n_t \cdot f_t(k) - n_t \cdot f_t(k) = (t + 1) \cdot f_{t+1}(k) - t \cdot f_t(k)$.

Số các nút bậc k tăng lên khi u nối với đỉnh v_i bậc $k - 1$ trong G_t , vì trong trường hợp này v_i có bậc k trong G_{t+1} . Theo q cạnh thêm vào tại bước $t + 1$, số các nút với bậc $k - 1$ trong G_t được chọn để nối đến u là $q \pi_t(k - 1) = q \cdot (k - 1) \cdot f_t(k - 1) / (2q) = (k - 1) \cdot f_t(k - 1) / 2$. Công thức này với giả thiết $k > q$, do v_i cần có bậc ít nhất là q ; mỗi nút được

¹ Master equation: phương trình chủ.

bổ sung tại bước $t \geq 1$ có bậc khởi đầu q . Do vậy, nếu $d_t = k - 1$, thì $k - 1 \geq q$, ngầm định $k > q$.

Đồng thời số các nút bậc k giảm khi nút u nối với đỉnh v_t bậc k trong G_t , vì trong trường hợp này v_t sẽ có bậc $k + 1$ trong G_{t+1} . Sử dụng công thức xác suất trên q cạnh thêm vào tại bước $t + 1$, số nút với bậc k trong G_t được chọn nối với u là $q^* \pi_t(k) = q^* k^* f_t(k) / (2q) = k^* f_t(k) / 2$. Dựa vào trao đổi trước, khi $k > q$, thay đổi mạng về số nút bậc k là sai khác giữa lượng tăng và giảm $q^* \pi_t(k - 1) - q^* \pi_t(k) = (k - 1)^* f_t(k - 1) / 2 - k^* f_t(k) / 2$. Từ đó rút ra phương trình chủ đối với $k > q$: $(t + 1)^* f_{t+1}(k) - t^* f_t(k) = (k - 1)^* f_t(k - 1) / 2 - k^* f_t(k) / 2$.

Mặt khác, khi $k = q$, giả sử không có nút đồ thị bậc thấp hơn q , thì chỉ nút mới bổ sung làm tăng số nút bậc $k = 1$ thêm một. Tuy nhiên, nếu u nối với nút v_t đã biết bậc k , thì sẽ giảm số nút bậc k , do lúc này v_t sẽ có bậc $k + 1$ trong G_{t+1} . Vậy, thay đổi mạng về số nút bậc k là: $1 - q^* \pi_t(k) = 1 - k^* f_t(k) / 2$. Vậy nên phương trình chủ khi $k = q$ là $(t + 1)^* f_{t+1}(k) - t^* f_t(k) = 1 - k^* f_t(k) / 2$. Dích chung là ổn định hay nghiệm không thay đổi theo thời gian được phương trình chủ. Nói khác đi, người ta nghiên cứu nghiệm khi $f_{t+1}(k) = f_t(k) = f(k)$. Nghiệm ổn định cho phân phối bậc độc lập với thời gian.

Trước hết suy diễn nghiệm ổn định đối với $k = q$. Thay $f_{t+1}(k) = f_t(k) = f(k)$ vào $(t + 1)^* f_{t+1}(k) - t^* f_t(k) = 1 - k^* f_t(k) / 2$, được $(t + 1)^* f(q) - t^* f(q) = 1 - q^* f(q) / 2$, nên $f(q) = 2 / (q + 2)$.

Nghiệm ổn định với $k > q$ cần tính $f(k)$ đệ quy theo $f(k - 1)$: $(t + 1)^* f(k) - t^* f(k) = (k - 1)^* f(k - 1) / 2 - k^* f(k) / 2$, nên $f(k) = ((k - 1) / (k + 2))^* f(k - 1)$. Mở rộng công thức này đến biên điều kiện $k = q$, thu được $f(k) = ((k - 1) / (k + 2))^* f(k - 1) = ((k - 1) (k - 2) / ((k + 2) (k + 1)))^* f(k - 1) = \dots = ((q + 2) (q + 1) q / ((k + 2) (k + 1) k))^* f(q)$. Thế kết quả $f(q)$ đã tính vào công thức, được $f(k) = ((q + 2) (q + 1) q / ((k + 2) (k + 1) k))^* (2 / (q + 2)) = 2q (q + 1) / (k (k + 1) (k + 2))$.

Với q không đổi và k lớn, phân phối bậc tỉ lệ theo $f(k) \propto k^{-\gamma}$. Nói cách khác, mô hình BA đạt đến phân phối bậc luật lũy thừa với $\gamma = 3$, đặc biệt khi bậc lớn.

Phân phối bậc theo tiếp cận liên tục: tiếp cận còn được gọi là phương pháp *trường trung bình*¹. Trong mô hình BA, các đỉnh được thêm vào sớm có khuynh hướng có bậc cao, do chúng có nhiều cơ hội lấy được liên kết từ các đỉnh sẽ thêm sau này. Phụ thuộc thời gian của bậc của đỉnh có thể xấp xỉ như biến ngẫu nhiên liên tục. Kí hiệu $k_i = d_i(t)$ là bậc của đỉnh v_i tại thời điểm t . Tại thời điểm t , xác suất nút u mới bổ sung liên kết đến v_i là $q \cdot \pi_i(t)$. Ngoài ra, thay đổi về bậc của v_i theo các bước thời gian là $q \cdot \pi_i(t)$. Sử dụng xấp xỉ $n_i \approx t$ và $m_i \approx q \cdot t$, tỉ lệ thay đổi của k_i theo thời gian được viết là: $\frac{dk_i}{dt} = q \cdot \pi_i(t) = q \cdot k_i / (2q \cdot t) = k_i / (2t)$. Lấy tích phân công

thức này, được $\int \frac{1}{k_i} dk_i = \int \frac{1}{2t} dt$, tức $\ln k_i = \ln t / 2 + C$, nên $k_i = \alpha \cdot \sqrt{t}$;

trong đó C là hằng số tích phân, nên $\alpha = e^C$ cũng là hằng số.

Kí hiệu t_i là thời gian khi nút i được bổ sung vào mạng. Do bậc khởi đầu với mọi nút là q , điều kiện biên là $k_i = q$ tại $t = t_i$. Đẳng thức trên với điều kiện biên: $k_i = \alpha \cdot \sqrt{t_i} = q$, nên $\alpha = q / \sqrt{t_i}$.

Thay vào biểu thức trên, $k_i = \alpha \cdot \sqrt{t} = q \cdot \sqrt{\frac{t}{t_i}}$.

Về trực quan, nghiệm này khẳng định hiện tượng nước chảy chỗ trũng. Nó gợi ý rằng nếu nút v_i bổ sung sớm vào mạng, tức t_i nhỏ, thì quá trình thời gian, tức t lớn lên, bậc của v_i tăng, tỉ lệ với căn của thời gian t .

Xét xác suất để bậc của v_i tại thời điểm t là nhỏ hơn giá trị k nào đó, tức $P(k_i < k)$. Nếu $k_i < k$, thì $q \cdot \sqrt{\frac{t}{t_i}} < k$, nên $t_i > q^2 t / k^2$. Vậy có thể viết $P(k_i < k) = P(t_i > q^2 t / k^2) = 1 - P(t_i \leq q^2 t / k^2)$. Nói cách khác, xác suất để nút v_i có bậc nhỏ hơn k là như xác suất để tại t_i mà v_i gia nhập mạng là lớn hơn $q^2 t / k^2$.

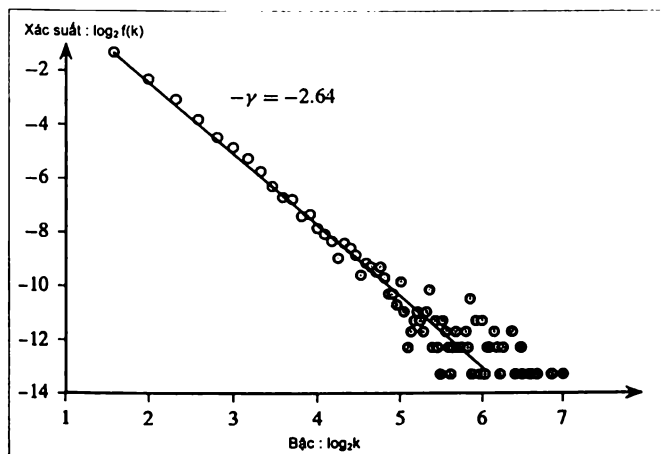
Các đỉnh thêm vào đồ thị theo tỉ lệ thống nhất là một nút trong một bước thời gian, tức $1/n_i \approx 1/t$. Vậy xác suất để $t_i \leq q^2 t / k^2$

¹ Mean field: trường trung bình.

là: $P(k_i < k) = 1 - P(t_i \leq q^2 t / k^2) = 1 - (q^2 t / k^2) * (1/t) = 1 - q^2 / k^2$. Do v_i là nút sinh bất kì trong đồ thị, $P(k_i < k)$ có thể xem như phân phối bậc tích lũy $F_t(k)$ tại thời điểm t . Người ta có thể có phân phối bậc $f_t(k)$ nhờ đạo hàm của $F_t(k)$ theo k : $f_t(k) = \frac{d}{dk} F_t(k) = \frac{d}{dk} P(k_i < k) = \frac{d}{dk} (1 - \frac{q^2}{k^2}) = 0 - ((k^2 * 0 - q^2 * 2k) / k^4) = \frac{2q^2}{k^3} \propto k^{-3}$.

Trong đẳng thức này, sử dụng $f(k) = g(k) / h(k)$, được $\frac{df(k)}{dk} = \frac{h(k) * \frac{dg(k)}{dk} - g(k) * \frac{dh(k)}{dk}}{h(k)^2}$. Ở đây, $g(k) = q^2$ và $h(k) = k^2$, và $\frac{dg(k)}{dk} = 0$ và $\frac{dh(k)}{dk} = 2k$.

Nhận xét thấy phân phối bậc theo tiếp cận liên tục là gần với tiếp cận rời rạc. Cả hai nghiệm khẳng định phân phối bậc tỉ lệ với k^{-3} , ứng với hành vi luật lũy thừa với $\gamma = 3$.



Hình 4.28. Mô hình Barabási - Albert, $n_0 = 3$, $t = 997$, $q = 3$: phân phối bậc

Hệ số phân cụm và đường kính: người ta cho biết đường kính của đồ thị BA tỉ lệ theo $d(G_t) = O(\log n_t / \log \log n_t)$, gợi ý rằng chúng thể hiện hành vi thế giới cực nhỏ, khi $q > 1$. Ngoài ra, hệ số

phân cụm kì vọng của đồ thị BA tỉ lệ theo $E[C(G_t)] = O((\log n)^2/n)$, chỉ tốt hơn một chút so với hệ số phân cụm đối với đồ thị ngẫu nhiên, mà tỉ lệ theo $O(n^{-1})$. Thí dụ sau nghiên cứu thực nghiệm hệ số phân cụm và đường kính đối với các biểu hiện ngẫu nhiên của mô hình BA, với tập tham số xác định.

Hình vẽ thể hiện phân phối bậc thực nghiệm từ trung bình của 10 đồ thị BA khác nhau, với tham số $n_0 = 3$, $q = 3$ và $t = 997$ bước thời gian, để đồ thị cuối cùng có $n = 1000$ đỉnh. Độ dốc của đường theo tỉ lệ $\log - \log$ khẳng định có luật lũy thừa, với độ dốc $\gamma = -2.64$.

Hệ số phân cụm trung bình trên 10 đồ thị và $C(G) = 0.019$, không quá cao, cho biết mô hình BA không nắm giữ hiệu ứng phân cụm. Mặt khác, đường kính trung bình là $d(G) = 6$, cho biết hành vi thể giới cực nhỏ.

4.4. Dữ liệu chuỗi thời gian

4.4.1. Giới thiệu về chuỗi thời gian

Định nghĩa: Một chuỗi thời gian có thể được coi như một tập hợp các giá trị, $\{x_t\}$, biểu diễn các phép đo được thực hiện tại các khoảng thời gian, tuần tự, khác nhau.

Thời gian $t = 1, 2, 3, \dots, n$, là biến liên tục, nhưng đo trong nhiều trường hợp được thực hiện tại các điểm cụ thể trong thời gian, vì vậy xuất hiện như những quan sát rời rạc. Thông thường một biến hoặc giá trị đơn lẻ được phân tích tại mỗi điểm trong thời gian, tức phân tích đơn biến, ngay cả khi nhiều biến được ghi nhận tại điểm đó. Chẳng hạn (i) điều kiện thời tiết hàng ngày tại trạm khí tượng; (ii) giá cổ phiếu tài chính theo giờ; (iii) khối lượng giao dịch dữ liệu. Thông thường, nhưng không nhất thiết, dữ liệu được đo hoặc được xác định tại các thời điểm có khoảng cách bằng nhau. Chẳng hạn sau mỗi phút, mỗi ngày, mỗi năm. Có thể phân tích chuỗi thời gian và xử lí nhiều biến số, nhưng không được đề cập chi tiết ở đây.

Khi nói về *tự tương quan thời gian*, có một số chuỗi thời gian (i) dữ liệu về giá cổ phiếu; (ii) dữ liệu về nhiệt độ. Người ta có cách phân tích dữ liệu để xác định các mẫu với *hàm tự tương quan*. Ở đây, tiếp tục phân tích thêm, xem xét các phương pháp khác để phân tích các mẫu trong dữ liệu loại này; từ đó, thiết kế các mô hình có chức năng giải thích và dự đoán. Bắt đầu bằng cách xem xét các phương pháp thống kê được áp dụng cho chuỗi thời gian trong miền thời gian, tức là phân tích dữ liệu bằng cách kiểm tra hành vi của chúng theo thời gian. Sau đó, xem xét những vấn đề tương tự, trong lĩnh vực tần số; các phân tích như vậy đặc biệt áp dụng cho dữ liệu thể hiện nhiều lần theo thời gian, có tính chu kì, và phân tích và mô hình hóa các khoảng thời gian này.

Chatfield, 1975, đã xác định một loạt các loại chuỗi thời gian khác nhau:

- *Chuỗi thời gian kinh tế*, như giá cổ phiếu, số liệu kinh tế vĩ mô, số liệu thu nhập;
- *Chuỗi thời gian vật lý*, như dữ liệu lưu lượng sông, dữ liệu khí tượng, dữ liệu giám sát ô nhiễm;
- *Chuỗi thời gian tiếp thị*, như số liệu doanh thu, số liệu phản hồi quảng cáo;
- *Chuỗi thời gian nhân khẩu học*, như mức dân số theo thời gian;
- *Quá trình sản xuất*, như bảng kết xuất và kiểm soát quá trình;
- *Quy trình nhị phân*, như các dãy số dữ liệu trong các hệ thống chuyển mạch và các hệ thống truyền tải dữ liệu;
- *Quá trình các điểm thời gian*, như dãy các điểm một chiều, hai chiều, dữ liệu không gian.

Đối với một số chuỗi dữ liệu, việc đo là liên tục, chẳng hạn đo áp suất không khí, hoặc kênh truyền dữ liệu qua đó lưu lượng được theo dõi liên tục. Các dữ liệu khác chỉ được đo hoặc xác định vào những thời điểm cụ thể, chẳng hạn giá niêm yết tại sàn giao dịch. Một số lớn các kĩ thuật thống kê mô tả vấn đề phân tích

chuỗi thời gian đối với loại dữ liệu loại thứ hai, tức dữ liệu rời rạc, thường được ghi lại theo khoảng cố định.

Có một số lí do mà các nhà thống kê thực hiện phân tích chuỗi thời gian. Các yêu cầu này bao gồm:

- Các yêu cầu mô tả đơn giản; chẳng hạn sử dụng các đồ thị và các biện pháp miêu tả để xác định các khía cạnh chính của dữ liệu, tức các đặc trưng trung bình, đỉnh dữ liệu, tính tuần hoàn... Cũng có thể là một yêu cầu sử dụng thông tin để dự đoán, ước lượng giá trị tương lai của chuỗi dữ liệu, dựa trên dữ liệu lịch sử, tức dữ liệu quá khứ;
- Dự đoán là một quy trình tương đối tĩnh, nhưng mô hình dự đoán thời gian thực ngày càng được sử dụng, đặc biệt là trong các môi trường năng động cao như kiểm soát khẩn cấp, quản lí cơ sở hạ tầng về truyền thông, điện lưới... và trong cả thị trường tài chính. Dự đoán chuỗi thời gian thuần nhất sử dụng dữ liệu thô để đưa ra dự đoán. Trong một số trường hợp, đặc biệt đối với dự đoán ngắn hạn, đây có thể là cách tiếp cận hiệu quả, vì dữ liệu phức tạp hơn, ứng với dữ liệu có hành vi phản ứng phức tạp, và dự đoán dài hạn, phân tích thống kê là rất quan trọng. Trong trường hợp này, dữ liệu về nhiều biến số có thể được sử dụng để cung cấp các ước lượng về các giá trị trong tương lai của chuỗi thời gian. Chẳng hạn dự đoán tổng sản phẩm quốc nội GDP¹, người ta thấy chuỗi thời gian là phức tạp, phụ thuộc vào nhiều biến số quốc gia và quốc tế;
- Mục tiêu phân tích chuỗi thời gian có thể là cho phép hoặc trợ giúp kiểm soát một số quá trình năng động. Chẳng hạn, các kĩ sư dự đoán nhu cầu điện cao điểm trong ngày dựa trên nhiều đầu vào nội bộ và bên ngoài và sau đó sử dụng những dự đoán này để tăng công suất phát điện trên

¹ GDP: Gross Domestic Product: tổng sản phẩm quốc nội.

dòng hoặc các nguồn cung trước các đỉnh cao về nhu cầu và sau đó giảm công suất khi yêu cầu giảm.

Chuỗi thời gian hầu như luôn biểu lộ mức *tự tương quan*, vì thế phân tích cấu trúc tự tương quan trong một chuỗi thường là một trong những nhiệm vụ đầu tiên được thực hiện, sau khi kiểm tra trực quan và làm sạch dữ liệu. Tuy nhiên, nhiều chuỗi thời gian cũng có một số đặc điểm chung khác. Chẳng hạn, nhiều chuỗi thời gian có tính chu kì, và chiều dài và cường độ của các mẫu ấy cần được kiểm tra chặt chẽ. Tương tự, chuỗi thời gian có thể hoặc không thể hiện thị *xu thế* riêng biệt trong dữ liệu, nhưng có lẽ một chuỗi dài hơn sẽ hiển thị xu thế khi không có mẫu nào ngắn hơn, do đó các điểm thời gian bắt đầu, kết thúc và thời lượng của chuỗi dữ liệu là rất quan trọng. Thường thì các xu thế rõ ràng cần được xác định và loại bỏ, trước khi thực hiện phân tích tự tương quan.

Định nghĩa: Chuỗi thời gian tĩnh là chuỗi mà thời điểm chuỗi được kiểm tra không ảnh hưởng đến kết quả, trong phân tích.

Hình thức hơn, một chuỗi thời gian tĩnh là một trong những phân phối xác suất chung không bị thay đổi bởi sự thay đổi về thời gian hoặc không gian; điều này đòi hỏi trung bình và phương sai của dữ liệu là không đổi theo thời gian, không gian. Nếu chuỗi có xu thế hay hành vi định kỳ, thường thì các thành phần này sẽ được nhận dạng và loại bỏ, tức chuỗi được *phân rã*, trước khi phân tích. Nhiều mô hình được sử dụng trong phân tích chuỗi thời gian, bao gồm (i) mô hình tự hồi quy đơn giản AR^1 ; (ii) mô hình trung bình động MA ; (iii) mô hình $ARMA^2$ kết hợp, giả định tính dừng. Các mô hình phức tạp hơn, chẳng hạn như $ARCH$ và $GARCH$, sử dụng trong mô hình kinh tế lượng.

¹ AR : Autoregressive: tự hồi quy; MA : Moving Average: trung bình động.

² $ARMA$: autoregressive moving average: trung bình động tự hồi quy.

4.4.2. Trung bình động

Với các bộ dữ liệu thông thường, giá trị trung bình thường là đặc trưng đầu tiên và là một trong những số liệu thống kê tóm tắt hữu ích nhất để tính toán. Khi dữ liệu ở dạng chuỗi thời gian, trung bình chuỗi là một phép đo hữu ích, nhưng không phản ánh tính năng động của dữ liệu. Các giá trị trung bình được tính trên các giai đoạn ngắn, hoặc trước giai đoạn hiện tại hoặc tập trung vào giai đoạn hiện tại, thường hữu ích hơn. Bởi vì các giá trị trung bình như vậy sẽ thay đổi, hoặc di chuyển. Chẳng hạn xét thời gian hiện tại, hoặc di chuyển từ thời gian $t = 2, t = 3 \dots$ Chúng được gọi là trung bình động, hay trung bình di động MA.

Định nghĩa: Trung bình động đơn giản là trung bình không trọng số của k giá trị đầu tiên.

Trung bình động có trọng số theo số mũ về cơ bản giống với trung bình động đơn giản, nhưng với tham gia của các trọng số. Do có nhiều chuỗi các giá trị trung bình động cho chuỗi thời gian, người ta vẽ trên đồ thị các MA, để phân tích dưới dạng chuỗi, và sử dụng trong mô hình hóa và dự đoán. Một loạt các mô hình có thể được xây dựng nhờ sử dụng trung bình động, và chúng được gọi là mô hình MA. Nếu các mô hình như vậy kết hợp với các mô hình tự hồi quy AR, các mô hình kết hợp kết quả được gọi là các mô hình ARMA hoặc ARIMA.

4.4.2.1. Trung bình động đơn giản

Vì một chuỗi thời gian có thể được coi như một tập các giá trị $\{x_t\}$, $t = 1..n$. Trung bình của các giá trị này là $\bar{x} = \sum_{t=1}^n x_t$. Nếu giả sử rằng n là khá lớn, và chúng ta chọn một số nguyên k nhỏ hơn n , có thể tính một tập hợp các trung bình khối, hoặc trung bình động đơn giản, của bậc k : $\bar{x}_{t,1} = \frac{1}{k} \sum_{t=1}^k x_t \dots \bar{x}_{t,(n-k+1)} = \frac{1}{k} \sum_{t=n-k+1}^n x_t$.

Mỗi phép đo đại diện cho mức trung bình của các giá trị dữ liệu trong một khoảng thời gian quan sát k . Thấy rằng MA đầu

tiền của bậc $k > 0$ là với $t = k$. Nói chung, có thể viết lại là $\hat{\mu}_t = \bar{x}_t = \frac{1}{k} \sum_{i=t-k+1}^t x_i$. Điều này cho biết trung bình ước lượng tại thời điểm t là trung bình đơn giản của giá trị quan sát được tại thời điểm t và các bước thời gian $k - 1$ trước đó. Nếu trọng số được áp dụng, tức làm giảm sự đóng góp của các quan sát xa hơn so với thời điểm hiện tại, trung bình động được làm *mượt theo số mũ*. Mức trung bình động thường được sử dụng như là một hình thức dự đoán, tức giá trị ước lượng cho chuỗi tại thời điểm $t + 1$, được S_{t+1} được lấy làm MA trong giai đoạn lên đến và bao gồm cả thời gian t .

Mức trung bình động đơn giản có thể được xem như là một hình thức làm mịn, loại bỏ một số thành phần tần số cao của một chuỗi thời gian và làm nổi bật, nhưng không loại bỏ, các xu thế, theo cách tương tự về lọc kĩ thuật số. Trung bình động là một dạng của bộ lọc tuyến tính. Có thể áp dụng tính toán trung bình động cho một chuỗi đã được làm trơn, tức làm trơn hoặc lọc chuỗi đã trơn. Chẳng hạn, với trung bình động của trật tự 2, người ta có thể xem nó như được tính bằng trọng số $\{1/2, 1/2\}$, do đó MA tại $x_2 = 0.5x_1 + 0.5x_2$. Tương tự, MA tại $x_3 = 0.5x_2 + 0.5x_3$. Nếu người ta áp dụng mức độ làm mịn hoặc lọc thứ hai, người ta có $0.5x_2 + 0.5x_3 = 0.5(0.5x_1 + 0.5x_2) + 0.5(0.5x_2 + 0.5x_3) = 0.25x_1 + 0.5x_2 + 0.25x_3$ tức là quy trình lọc hai giai đoạn, hay xoắn, đã tạo ra trung bình động đối xứng, có trọng số, theo $\{1/4, 1/2, 1/4\}$. Nhiều lần xoắn có thể tạo ra trung bình động có trọng số phức tạp, như thấy trong ứng dụng đặc biệt trong các lĩnh vực chuyên ngành, chẳng hạn như tính toán bảo hiểm nhân thọ.

Trung bình động có thể dùng để loại bỏ các hiệu ứng chu kì, nếu tính với thời lượng của chu kỳ. Chẳng hạn, với dữ liệu hàng tháng, biến thiên theo mùa thường bị bỏ qua, theo cách sử dụng trung bình động 12 tháng đối xứng với tất cả các tháng cùng trọng số, trừ tháng đầu và tháng cuối có trọng số $1/2$. Làm như vậy vì sẽ có 13 tháng trong mô hình đối xứng, tức thời điểm hiện tại $t \pm 6$

tháng. Tổng cộng được chia cho 12. Có thể áp dụng các thủ tục tương tự cho bất kỳ tính chu kì, định kỳ nào.

4.4.2.2. Trung bình động có trọng số theo số mũ

Trong công thức về trung bình động đơn giản $\hat{\mu}_t = \bar{x}_t = \frac{1}{k} \sum_{i=t-k+1}^t x_i$, các quan sát không đánh trọng số, tức trọng số như nhau. Nếu kí hiệu trọng số α_i thì $\alpha_i = 1/k$, và tổng các trọng số là 1. Tức công thức có dạng $\hat{\mu}_t = \bar{x}_t = \sum_{i=t-k+1}^t \alpha_i x_i$.

Người ta thấy rằng nhiều ứng dụng của quá trình này theo các trọng số khác nhau. Với *trung bình động có trọng số theo số mũ* EWMA¹, sự đóng góp vào giá trị trung bình từ các quan sát ít có giá trị thời gian sẽ được xem xét giảm, do đó nhấn mạnh các sự kiện gần nhất, cục bộ. Về cơ bản một tham số làm mịn, $0 < \alpha < 1$, được sử dụng, thu được $\hat{\mu}_t = \bar{x}_t = \sum_{i=t-k+1}^t \alpha_i x_i$, trong đó $\sum_{i=t-k+1}^t \alpha_i = 1$. Phiên bản đối xứng của công thức này là $\hat{\mu}_t = \bar{x}_t = \sum_{i=t-q}^{t+q} \alpha_i x_i$, trong đó $\sum_{i=t-q}^{t+q} \alpha_i = 1$. Nếu trọng số trong mô hình đối xứng được chọn là các hạng thức mở rộng nhị thức, $(1/2 + 1/2)^{2q}$, chúng có tổng là 1, và khi q trở nên lớn, sẽ dẫn đến *phân phối chuẩn*. Đây là một dạng phân phối nhân, với *nhị thức* hoạt động như hàm nhân. Việc ghép hai giai đoạn mô tả trong phần trước đây, chính xác là sắp đặt này, với $q = 1$, đạt đến trọng số $\{1/4, 1/2, 1/4\}$.

Trong việc làm mịn số mũ, cần sử dụng một bộ trọng số có tổng 1 và làm giảm kích thước về hình học. Trọng số được sử dụng thường có dạng $\alpha(1-\alpha)^i$, $i = 0.. \infty$. Theo công thức nhị thức và triển khai công thức, thu được $\sum_{i=0}^{\infty} \alpha(1-\alpha)^i = 1$. Vậy viết lại được $\hat{\mu}_t = \sum_{i=1}^t \alpha(1-\alpha)^{i-1} x_{t-i+1}$. Viết theo đệ quy, được công thức $\hat{\mu}_t = \alpha x_t + (1-\alpha)\hat{\mu}_{t-1}$, đơn giản hoá tính toán rất nhiều, và tránh được

¹ EWMA: Exponentially weighted moving average: trung bình động có trọng số theo số mũ.

vấn đề mà chế độ trọng số phải là vô hạn để tổng trọng số là 1. Một số người sử dụng S để cho thấy công thức cơ bản là một biến được làm mịn, và viết $S_t = \alpha x_t + (1 - \alpha)S_{t-1}$, trong khi lí thuyết điều khiển thường sử dụng Z thay vì S cho các giá trị trọng số mũ hoặc làm giá trị mịn số học.

Dạng $S_t = \alpha x_t + (1 - \alpha)S_{t-1}$ có thể phù hợp hơn trong một số quy trình điều khiển. Với $\alpha = 1$ ước lượng trung bình chỉ đơn giản là giá trị đo. Với $\alpha = 0.5$ ước lượng là mức trung bình động đơn giản của các phép đo hiện tại và trước đó. Trong các mô hình dự đoán, giá trị S_t thường được sử dụng làm giá trị dự đoán hoặc giá trị ước lượng trong khoảng thời gian tiếp theo, tức ước lượng cho x tại thời điểm. Vậy người ta có $\hat{x}_{t+1} = \alpha x_t + (1 - \alpha)\hat{\mu}_{t-1} = \alpha x_t + \hat{\mu}_{t-1} - \alpha\hat{\mu}_{t-1} = \alpha(x_t - \hat{\mu}_{t-1}) + \hat{\mu}_{t-1}$. Vậy nên $\hat{x}_{t+1} = \alpha(x_t - \hat{x}_t) + \hat{\mu}_{t-1} = \alpha \varepsilon_t + \hat{\mu}_{t-1}$. Điều này cho thấy giá trị dự đoán tại thời điểm $t + 1$ là sự kết hợp của trung bình động trọng số theo số mũ cộng với một thành phần đại diện cho lỗi dự đoán trọng số, tại thời điểm t .

Giả sử có một chuỗi thời gian, cần dự đoán, người ta cần xác định giá trị α . Nó có thể được ước lượng từ các dữ liệu hiện có bằng cách đánh giá tổng của các sai số bình phương của dự đoán, với các giá trị thời điểm khác nhau, $t = 2, 3, \dots$ và lấy ước lượng đầu tiên là giá trị dữ liệu quan sát đầu tiên, tức x_1 . Trong các ứng dụng điều khiển, giá trị α là quan trọng ở chỗ nó được sử dụng để xác định các giới hạn điều khiển trên và dưới, và tác động đến chiều dài chạy trung bình ARL¹ dự kiến trước khi các giới hạn điều khiển bị phá vỡ, với giả định rằng chuỗi thời gian thể hiện tập các biến độc lập phân phối ngẫu nhiên, có phương sai chung. Dưới các điều kiện, theo Lucas và Saccucci, năm 1990, biến thiên của thống kê điều khiển $S_t = \alpha x_t + (1 - \alpha)S_{t-1}$ là $\text{var}(S_t) = \alpha \sigma_x^2 \frac{1 - (1 - \alpha)^{2t}}{2 - \alpha}$; khi t

tăng và α không đủ lớn $\text{var}(S_t) \rightarrow \sigma_x^2 \frac{\alpha}{2 - \alpha}$.

¹ ARL: average run length: độ dài chạy trung bình.

Giới hạn điều khiển thường được đặt thành bội số cố định của phương sai tiệm cận, chẳng hạn ± 3 lần độ lệch chuẩn. Nếu $\alpha = 0.25$ và dữ liệu đang được giám sát giả định có phân phối chuẩn, $N(0, 1)$, khi trong điều khiển, các giới hạn kiểm soát sẽ là ± 1.134 và quá trình sẽ đạt đến giới hạn sau trung bình 500 bước. Lucas và Saccucci, năm 1990, rút ra ARL đối với phạm vi rộng các giá trị và tuân theo các giả thiết về quy trình xích Markov. Họ lập bảng kết quả, bao gồm ARL khi trung bình của quá trình điều khiển đã được trượt đi vài độ lệch chuẩn. Chẳng hạn, với trượt 0.5 với $\alpha = 0.25$, thì ARL ít hơn 50 bước thời gian.

Các tiếp cận mô tả ở trên được gọi là *làm mịn mũ đơn*, vì các quy trình được áp dụng một lần cho chuỗi thời gian và sau đó các quy trình phân tích hoặc điều khiển được thực hiện trên bộ dữ liệu được làm mịn. Nếu bộ dữ liệu bao gồm xu thế hay thành phần theo mùa, việc mũ hóa theo hai hoặc ba giai đoạn có thể được áp dụng trung bình động như cách loại bỏ những ảnh hưởng này, tức là mô hình tường minh.

4.4.3. Phân tích xu thế

Khi các chuỗi thời gian có các hành vi xu thế hay có tính chu kỳ, thông thường các thành phần này được xác định và loại bỏ trước khi phân tích tiếp. Phần này sẽ xem xét các cách khác nhau để xác định và loại bỏ các xu thế và các hiệu ứng theo mùa, có tính chu kỳ, và các công cụ để tách một chuỗi thời gian thành *xu thế*, các thành phần *theo mùa* và các thành phần còn lại. Việc này được gọi là *phân rã*. Điều này dẫn đến việc xem xét các mô hình tiên tiến hơn, dựa trên những yếu tố này để tạo ra các mô hình phức hợp nhiều thành phần hơn, tạo cơ sở cho dự đoán thống kê.

4.4.3.1. Việc phân tích xu thế

Nhiều chuỗi thời gian có xu thế, tức là chúng không ổn định. Xu thế có thể là tuyến tính, hoặc có dạng phức tạp hơn, chẳng hạn như đa thức hoặc *logit*. Bất kể xu thế đó ra sao, thông thường

người ta loại bỏ chúng, trước khi (i) phân tích; (ii) mô hình hóa. Có một số cách để thực hiện việc này. Trước tiên thường là kiểm tra đồ thị của chuỗi thời gian một cách trực quan, để xem có bất kỳ hành vi xu thế nào có thể được quan sát hay không. Có thể có chuỗi có xu thế mà không dễ quan sát, mà chỉ hiện ra xu thế trên một dải dữ liệu, hay một phần của dải dữ liệu đó. Trong trường hợp này, người ta thường tách chuỗi thời gian thành một số bộ dữ liệu con; mô hình hóa riêng từng bộ dữ liệu con. Phân tích *tự hồi quy* là kĩ thuật rất hữu ích để xác định các khuynh hướng và chu kỳ của dữ liệu, theo cách thức thường chính xác hơn so với kiểm tra trực quan.

Sau khi xác định được một xu thế tồn tại, người ta có thể xem xét các thủ tục xác định và quản lí các xu thế. Các thủ tục đó là:

- *Đường cong phù hợp*. Chẳng hạn áp dụng hồi quy tuyến tính theo bình phương tối thiểu, hay đường cong tăng trưởng đối với dữ liệu; các đường cong tăng trưởng như hàm logit hay hàm Gompertz;
- *Lọc*. Người ta làm mịn một lần hay nhiều lần với lọc tuyến tính, hay trung bình động;
- *Tính sự khác biệt*. Sử dụng toán tử trừ đơn giản $\Delta t = x_t - x_{t-1}$, sinh ra chuỗi mới $\{y_t\}$ từ chuỗi $\{x_t\}$ hiện có; khi đó xu thế tuyến tính đã bị xóa.

Trong mỗi thủ tục này, mục tiêu là loại bỏ xu thế và tạo ra chuỗi thời gian cố định bao gồm các kết quả của hoạt động hoặc phần còn lại sau khi loại bỏ xu thế. Rồi bộ dữ liệu mới này được phân tích sâu hơn, chẳng hạn để mô hình hóa hành vi của các phần còn lại, và sau đó có thể kết hợp lại với các yếu tố xu thế để tạo ra mô hình mô tả các chuỗi thời gian và cung cấp khả năng dự đoán hành vi tương lai của nó.

Trong phần trước, trên các giá trị *trung bình động có trọng số theo số mũ EWMA*, đã chỉ ra rằng trung bình động tại thời điểm t có thể thu được như là tổng có trọng số của giá trị quan sát gần

nhất x_t và EWMA trước đó: $\hat{\mu}_t = \alpha x_t + (1 - \alpha)\hat{\mu}_{t-1}$, $0 \leq \alpha \leq 1$. Mô hình này không cho phép các xu thế trong dữ liệu. Nếu một chuỗi thời gian có xu thế tuyến tính đơn giản, và kí hiệu b_t đại diện cho tỉ lệ thay đổi dữ liệu trên một đơn vị thời gian, thì có thể tăng thêm biểu thức trước bằng cách đưa vào thành phần tỉ lệ $\hat{\mu}_{t-1} \rightarrow \hat{\mu}_{t-1} + b_{t-1}$, thu được quan hệ đệ quy chỉnh sửa $\hat{\mu}_t = \alpha x_t + (1 - \alpha)(\hat{\mu}_{t-1} + b_{t-1})$, hay theo kí pháp làm mịn thì $S_t = \alpha x_t + (1 - \alpha)(S_{t-1} + b_{t-1})$.

Điều này đặt ra vấn đề về việc làm thế nào để xác định tỉ lệ thay đổi, b_t , và nó nên là hằng số hay biến thay đổi theo thời gian. Thường người ta giả định tỉ lệ thay đổi theo thời gian và được ước lượng lặp theo: $b_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)b_{t-1}$, với $0 \leq \gamma \leq 1$. Điều này đảm bảo mô hình được làm mịn theo cấp mũ gấp đôi. Đơn giản nhất, khởi đầu với $S_1 = x_1$ và $b_1 = x_2 - x_1$. Các thông số được ước lượng thông thường bởi thủ tục tối ưu bình phương tối thiểu.

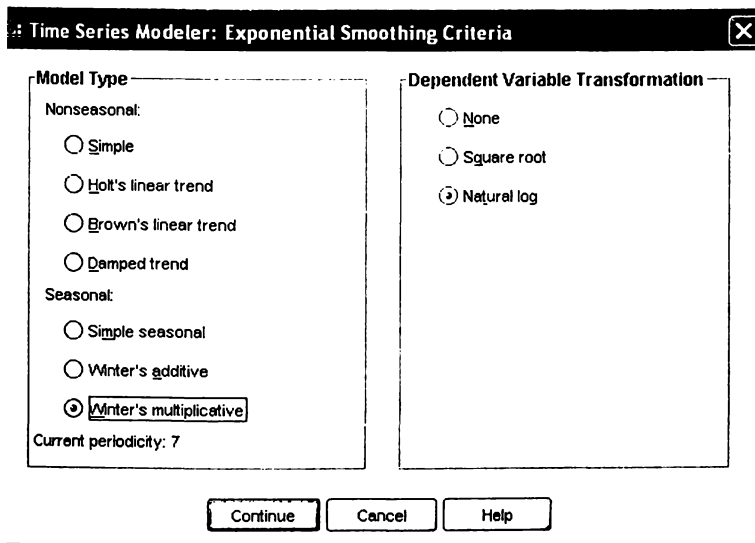
4.4.3.2. Phân tích theo mùa, định kỳ

Dữ liệu chuỗi thời gian thường biểu hiện tính chu kì, trong một số trường hợp phản ánh thời vụ, chẳng hạn như sự tăng trưởng thực vật hoặc trong các mẫu mua sắm. Các mô hình này có thể được xác định bằng cách (i) kiểm tra bằng thị giác trên đồ thị chuỗi thời gian; (ii) phân tích tự tương quan, chẳng hạn *phân tích phổ* theo phân tích trong miền tần số. Khi đã xác định được các mô hình đó, chúng có thể được loại bỏ, hoặc được xây dựng một cách rõ ràng thành các mô hình, để có thể được sử dụng cho mục đích mô tả và dự đoán.

Trung bình động đơn giản, chẳng hạn *trung bình động hướng tâm*, kéo dài thời gian hiện tại t thêm $\pm p/2$, trong đó p là khoảng thời gian được để cập, sẽ loại bỏ hiệu ứng theo mùa trong giai đoạn p . Trung bình động trong trường hợp này bao gồm $p + 1$ hạng thức, nên hạng thức đầu và hạng thức cuối sẽ lấy trọng số 0.5, và số chia là p . Chẳng hạn, với dữ liệu theo tháng dữ liệu theo mùa, người ta loại bỏ, hoặc làm mịn hay lọc, theo trung bình động có dạng:

$$\hat{y}_{t+12} = \frac{0.5 * x_{t-6} + x_{t-5} + \dots + x_{t-1} + x_t + x_{t+1} + \dots + x_{t+5} + 0.5 * x_{t-6}}{12}$$

Tính sai khác đơn giản cũng loại bỏ được hiệu ứng mùa, tức sử dụng toán tử $\Delta t = x_t - x_{t-1}$. Nếu các chuỗi thời gian có cả xu thế lẫn hiệu ứng theo mùa, và chúng được xác định tương đối tốt, có thể mở rộng các ý tưởng làm mượt theo số mũ, như mô tả ở trên, để làm *mịn mũ gấp ba lần*. Nó đòi hỏi ba phương trình (i), phương trình làm mịn theo số mũ ban đầu; (ii) phương trình xu thế; (iii) phương trình làm mịn theo mùa. Các phương trình có các tham số có giá trị tối ưu có dạng tính lặp. Thủ tục này được gọi là *thủ tục dự đoán Holt - Winters*¹.



Hình 4.29. Giao diện xử lý dữ liệu theo mùa trong phần mềm SPSS

Khi các thành phần theo mùa có trong các mô hình loại này, các hiệu ứng có thể có cơ sở phụ trội đơn giản, tức *tăng thêm* vào trung bình, hoặc *gấp bội*, trong đó hiệu ứng mùa tăng hay giảm

¹Holt, năm 1957, và Winters, năm 1960, đề xuất giải pháp làm mịn chuỗi dữ liệu có hiệu ứng mùa.

một số lần, chứ không chỉ một hằng số. Nếu nghi ngờ có hiệu ứng gấp bội, người ta sử dụng biến đổi log, để chuyển về hiệu ứng tăng thêm. Các gói phần mềm cung cấp trợ giúp phân tích chuỗi thời gian thường cung cấp một loạt các lựa chọn để xây dựng các mô hình loại này. Chẳng hạn, như hình minh họa, phần mềm SPSS¹ trợ giúp mô hình làm mịn theo số mũ, theo mùa, có và không có sự chuyển đổi dữ liệu, nhưng có xác định trước về dữ liệu chu kỳ.

4.4.3.3. *Phân rã*

Khi cho chuỗi thời gian, được cho là (i) có xu thế kết hợp các xu thế cục bộ; (ii) thành phần theo mùa; và (iii) thành phần còn lại, người ta dùng công cụ để tách các thành phần này thành một quá trình đơn giản. Thủ tục này được gọi là *phân rã* chuỗi thời gian, và có sẵn trong nhiều gói phần mềm. Người ta có thể đồng thời phân rã và vẽ các kết quả.

4.4.4. *Các mô hình ARMA và ARIMA (Box - Jenkins)*

Các giá trị của một chuỗi thời gian một biến tại thời điểm t, x_t có thể được mô phỏng khi dùng nhiều biểu thức *trung bình động*. Các thành phần như xu thế và chu kỳ trong chuỗi thời gian có thể được mô hình tương minh, và tách ra; dữ liệu được phân rã thành xu thế, các thành phần theo mùa và các thành phần còn lại. Ngoài ra, các hệ số tương quan toàn phần và hệ số tương quan một phần rất có ích để xác định và mô hình hoá các chuỗi thời gian. Hai khía cạnh (i) phân tích chuỗi thời gian; và (ii) mô hình hóa có thể được kết hợp trong một khuôn khổ mô hình tổng quát hơn và thường có hiệu quả cao. Theo dạng cơ bản, tiếp cận này được gọi là mô hình trung bình động tự hồi quy ARMA, hoặc khi lấy hiệu số trong thủ tục sẽ được mô hình ARIMA hoặc Box - Jenkins. Không

¹ SPSS: Statistical Package for the Social Sciences: *bộ phần mềm thống kê cho khoa học xã hội*, là bộ phần mềm xử lý thống kê.

có quy tắc cố định về số khoảng thời gian cần thiết để thực hành thành công việc mô hình hóa, nhưng với bài toán phức tạp, nên cần chuỗi với khoảng 50 bước thời gian.

Các mô hình ARMA kết hợp phương pháp tự tương quan, tức AR, và trung bình động, tức MA, thành một mô hình tổng hợp của chuỗi thời gian. Trước khi xem xét làm cách kết hợp các mô hình này, phần sau đề cập các mô hình lẻ.

4.4.4.1. Mô hình MA

Phần trên đã cho thấy mô hình trung bình động MA có thể được sử dụng để phù hợp với một số bộ dữ liệu và các biến thể trên các mô hình này, có liên quan đến việc làm mịn mũ gấp đôi hoặc gấp ba có thể xử lý các xu thế và các thành phần định kỳ trong dữ liệu. Hơn nữa, các mô hình như vậy có thể được sử dụng để tạo ra các dự đoán theo hành vi của giai đoạn trước đó. Một mẫu đơn giản của các mô hình như vậy, dựa trên dữ liệu trước, có thể được viết như sau: $\hat{x}_t = \beta_0 x_t + \beta_1 x_{t-1} + \dots + \beta_q x_{t-q}$, tức $\hat{x}_t = \sum_{i=0}^q \beta_i x_{t-i}$; trong đó các β_i là các trọng số được áp dụng cho các giá trị trước trong chuỗi thời gian, và thường là để xác định $\beta_1 = 1$ mà không mất tính tổng quát. Vì vậy, đối với quá trình thứ tự đầu $q = 1$, có mô hình: $\hat{x}_t = \beta_0 x_t + \beta_1 x_{t-1}$, hay $\hat{x}_t = x_t + \beta_1 x_{t-1}$, với $\beta_0 = 1$. Nó cho thấy giá trị trung bình động được ước lượng là trung bình trọng số của giá trị hiện tại và giá trị trước đó. Quá trình trung bình này, theo nghĩa nào đó, là một cơ chế làm mịn thực tế mà không liên kết trực tiếp với mô hình thống kê. Tuy nhiên, có thể chỉ định một mô hình thống kê, hay mô hình ngẫu nhiên, với các thủ tục trung bình động kết hợp với các quá trình ngẫu nhiên. Nếu kí hiệu $\{z_t\}$ là tập các biến ngẫu nhiên độc lập và phân phối ngẫu nhiên, tức được quá trình ngẫu nhiên, với trung bình 0 và phương sai cố định, thì có thể viết quá trình $\{x_t\}$ như trung bình động bậc q , theo dạng thức của $\{z_t\}$: $\hat{x}_t = \beta_0 z_t + \beta_1 z_{t-1} + \dots + \beta_q z_{t-q}$, hay $\hat{x}_t = \sum_{i=0}^q \beta_i z_{t-i}$.

Rõ ràng giá trị kì vọng của x_t theo mô hình này là 0, do đó mô hình chỉ được kiểm chứng nếu x_t đã được điều chỉnh để có trung

bình 0 hoặc nếu thêm một hằng số cố định, tức trung bình của x_t vào tổng. Phương sai của x_t đơn giản là: $\text{var}(x_t) = \text{var}(\beta_0 z_t) + \text{var}(\beta_1 z_{t-1}) + \dots + \text{var}(\beta_q z_{t-q})$, hay $\text{var}(x_t) = \sigma_z^2 \sum_{i=0}^q \beta_i^2$. Phân tích này có thể mở rộng để đánh giá hiệp phương sai, $\text{cov}(x_t, x_{t+k})$, thu được: $\gamma(k) = \text{cov}(x_t, x_{t+k}) = \sigma_z^2 \sum_{i=0}^q \beta_i \beta_{i+k}$. Cả giá trị trung bình, lẫn hiệp phương sai, hay *tự hiệp phương sai*, đều không là hàm của thời gian t , tại trễ k ; nên quá trình này là *tĩnh độ hai*. Biểu thức này cho phép người ta thu được *hàm tự tương quan* ACF¹: $\rho_k = \rho(k) = \frac{\gamma(k)}{\gamma(0)} = \sum_{i=0}^{q-k} \beta_i \beta_{i+k} / \sum_{i=0}^q \beta_i^2$, $k = 0..q$. Nếu $k = 0$, $\rho_k = 1$, và với $k > q$, $\rho_k = 0$. Ngoài ra, hàm ACF là đối xứng và $\rho_k = \rho \cdot k$. Có thể tính ACF đối với quá trình MA thứ tự đầu: $x_t = z_t + \beta z_{t-1}$ và $\rho_1 = \rho_{-1} = \frac{\beta^2}{1 + \beta^2}$, $\rho_0 = 1$, $\rho_k = 0$, $k > 1$.

4.4.4.2. Mô hình AR

Các thành phần tự hồi quy AR của một mô hình ARMA có thể được viết dưới dạng $x_t = \alpha_1 x_{t-1} + \dots + \alpha_p x_{t-p} + z_t$, hay $x_t = \sum_{i=1}^p \alpha_i x_{t-i} + z_t$; trong đó các hạng thức là các *hệ số tự tương quan* tại trễ 1, 2, ..., p và z_t là hạng thức sai số còn lại. Hạng thức sai số đặc biệt liên quan đến khoảng thời gian hiện tại, t . Vì vậy, đối với quá trình thứ tự đầu, người ta có mô hình: $x_t = \alpha x_{t-1} + z_t$, $x_{t-1} = \alpha x_{t-2} + z_{t-1}, \dots$

Các biểu thức này chỉ ra rằng giá trị ước lượng của x tại thời điểm t được xác định trực tiếp theo giá trị trước đó của x , tức tại thời điểm $t - 1$. Giá trị này nhân với độ đo α nhằm mở rộng về giá trị để tất cả các cặp giá trị tại các khoảng thời gian trễ 1 có tương quan, tức tự tương quan. Cần cộng thêm hạng thức sai số còn lại z tại t . Đó chính là định nghĩa của *quá trình Markov*, nên quá trình Markov là quá trình tự hồi quy bậc một. Nếu $\alpha = 1$, mô hình này khẳng định rằng giá trị tiếp theo của x chỉ đơn giản là giá trị trước

¹ ACF: Autocorrelation function: hàm tự tương quan.

cộng với sai số ngẫu nhiên, và do đó là *bước đi ngẫu nhiên*¹ đơn giản. Nếu có nhiều hạng thức hơn, mô hình sẽ ước lượng giá trị của x tại thời điểm t theo tổng có trọng số của các hạng thức, cộng với thành phần sai số ngẫu nhiên.

Nếu chúng ta thay thế x_{t-1} và x_t được $x_t = \alpha(\alpha x_{t-2} + z_{t-1}) + z_t$, tức $x_t = \alpha^2 x_{t-2} + \alpha z_{t-1} + z_t$, và lặp lại việc này: $x_t = \alpha^k x_{t-k} + \alpha^{k-1} z_{t-k+1} + \dots + \alpha z_{t-1} + z_t$, hay $x_t = \sum_{i=1}^k \alpha^i x_{t-i} + z_t$.

Bây giờ nếu $|\alpha| < 1$ và k là lớn, biểu thức này có thể được viết theo thứ tự ngược lại, với điều kiện giảm dần và với sự đóng góp từ thuật ngữ trong x ở phía bên tay phải của biểu thức trở nên nhỏ bé, vì vậy chúng ta có: $x_t = z_t + \alpha z_{t-1} + \alpha^2 z_{t-2} + \dots$, hay $x_t = \sum_{i=0}^{\infty} \alpha^i z_{t-i}$. Do phía phải của biểu thức này, x_t là tổng của tập có trọng số các giá trị trước, trong trường hợp này là các hạng thức sai số ngẫu nhiên, nên rõ ràng mô hình AR này trên thực tế là một dạng của mô hình MA. Và nếu chúng ta giả định rằng các hạng thức sai số có trung bình 0 và phương sai không đổi, thì như đã có trong mô hình MA, giá trị kì vọng của mô hình cũng là 0, với giả thiết x_t đã được điều chỉnh để có trung bình 0, với phương sai là $\text{var}(x_t) = \text{var}(z_t + \alpha z_{t-1} + \alpha^2 z_{t-2} + \dots) = \text{var}(z) + \text{var}(\alpha z) + \text{var}(\alpha^2 z) + \dots$, nên $\text{var}(x_t) = \text{var}(z)(1 + \alpha + \alpha^2 + \alpha^3 + \dots)$. Khi $|\alpha| < 1$, tổng này là hữu hạn, là $1/(1 - \alpha)$, nên cuối cùng: $\text{var}(x_t) = \frac{\sigma_z^2}{1 - \alpha} = \sigma_x^2$. Giống như mô hình MA ở trên, phân tích này có thể được mở rộng để đánh giá hiệp phương sai, $\text{cov}(x_t, x_{t-k})$, được $\gamma(k) = \text{cov}(x_t, x_{t-k}) = \sigma_z^2 \sum_{i=0}^{\infty} \alpha^i \alpha^{i+k}$, $k > 0$; nên $\rho_k = \frac{\gamma(k)}{\gamma(0)} = \alpha^k$.

Điều này chứng tỏ rằng đối với mô hình tự hồi quy bậc nhất thì hàm tự tương quan ACF được xác định đơn giản bởi các lũy thừa liên tiếp của tự tương quan đầu tiên α , với điều kiện $|\alpha| < 1$. Đối với $\alpha > 0$, đây chỉ đơn giản là lũy thừa giảm nhanh hoặc đường cong hàm mũ, tiến về 0, hoặc đối với $\alpha < 0$ nó là một đường

¹ Random walk: bước đi ngẫu nhiên.

cong dao động làm giảm, cũng dần về 0. Nếu có giả thiết về chuỗi thời gian là tĩnh thì phân tích trên có thể được mở rộng cho các tự tương quan bậc hai và bậc cao hơn.

Để làm mô hình AR khít với bộ dữ liệu quan sát, người ta cố gắng giảm thiểu tổng các sai số bình phương, tức phù hợp theo bình phương tối thiểu, sử dụng lượng nhỏ nhất các hạng thức để dữ liệu khít. Các mô hình loại này được mô tả là *tự hồi quy*, và có thể được áp dụng cho cả chuỗi thời gian và các bộ dữ liệu không gian. Mặc dù theo lý thuyết, mô hình tự hồi quy có thể phù hợp với một số liệu quan sát, nhưng thông thường nó đòi hỏi phải loại bỏ trước và các xu thế và các thành phần định kỳ, và thậm chí sau đó có thể cần một số lượng lớn các hạng thức để khít với dữ liệu. Tuy nhiên, bằng cách kết hợp các mô hình AR với mô hình MA, có thể cho ra một họ các mô hình hỗn hợp, cho phép áp dụng trong nhiều tình huống. Các mô hình này được gọi là mô hình ARMA và ARIMA.

4.4.4.3. Mô hình ARMA

Các phần trên đã đề cập (i) mô hình MA bậc q : $\hat{x}_t = \beta_0 z_t + \beta_1 z_{t-1} + \dots + \beta_q z_{t-q}$, hay $\hat{x}_t = \sum_{i=0}^q \beta_i z_{t-i}$; và (ii) mô hình AR bậc p : $x_t = \alpha^k x_{t-k} + \alpha^{k-1} z_{t-k+1} + \dots + \alpha z_{t-1} + z_t$, hay $x_t = \sum_{i=1}^k \alpha^i x_{t-i} + z_t$. Có thể kết hợp hai mô hình này bằng cách ghép chúng lại với nhau như mô hình bậc (p, q) , $x_t = \alpha^k x_{t-k} + \alpha^{k-1} z_{t-k+1} + \dots + \alpha z_{t-1} + z_t + \beta_1 z_{t-1} + \dots + \beta_q z_{t-q}$.

Nói chung, dạng mô hình ARMA kết hợp này có thể dùng để mô hình hóa chuỗi thời gian với ít hạng thức hơn hai mô hình thành phần. Nó thể hiện giá trị ước tính tại thời gian t là tổng q hạng thức, ứng với phương sai trung bình của thay đổi ngẫu nhiên trên q khoảng thời gian trước, tức thành phần MA, cộng với tổng của p hạng thức AR tính cho giá trị t hiện tại như là tổng có trọng số của p giá trị gần nhất.

Tuy nhiên, dạng mô hình này giả định rằng chuỗi thời gian là tĩnh tại, tuy hiếm khi như vậy. Thực tế cho thấy xu thế và chu kỳ có trong nhiều bộ dữ liệu, do đó cần phải loại bỏ các hiệu ứng này, trước khi áp dụng các mô hình. Việc loại bỏ thường được thực hiện nhờ lấy hiệu số trong mô hình, một, hai lần, cho đến khi chuỗi là xấp xỉ tĩnh. Như với quá trình MA và AR, việc lấy hiệu số được mô tả theo thứ tự của việc lấy hiệu số, chẳng hạn $d = 1, 2, 3...$ Bộ ba (p, d, q) xác định dạng mô hình áp dụng. Mô hình này được mô tả như mô hình ARIMA¹.

4.4.4.4. Mô hình ARIMA

Trong thống kê và kinh tế lượng, và đặc biệt trong phân tích chuỗi thời gian, mô hình *trung bình động tích hợp tự hồi quy* ARIMA² là tổng quát của mô hình trung bình động tự hồi quy ARMA. Cả hai mô hình này đều phù hợp với dữ liệu chuỗi thời gian, để (i) hiểu rõ hơn về dữ liệu; (ii) dự đoán các điểm tương lai trong chuỗi. Các mô hình ARIMA được áp dụng trong một số trường hợp khi dữ liệu có dấu hiệu không ổn định, trong đó một bước phân biệt ban đầu, ứng với *phân tích hợp* của mô hình, có thể được áp dụng một hoặc nhiều lần để loại bỏ sự không ổn định.

Phần AR của ARIMA chỉ ra rằng biến số biến đổi quan tâm được hồi quy dựa trên giá trị trễ, tức tụt hậu, trước đây, của chính nó. Phần MA chỉ ra rằng sai số hồi quy thực sự là tổ hợp tuyến tính của hạng thức về sai số, tức giá trị xảy ra hiện tại và tại thời điểm khác nhau trong quá khứ. Các I, tức *tích hợp*, chỉ ra rằng các giá trị dữ liệu đã được thay thế bằng hiệu giữa các giá trị của chúng và các giá trị trước đó, và là quá trình lấy hiệu số này có thể đã được thực hiện nhiều lần. Mục đích của mỗi tính năng này là làm cho mô hình phù hợp với dữ liệu hơn.

¹ Chữ I với nghĩa tích hợp, hay với nghĩa căn tính hiệu số trên bộ dữ liệu ban đầu.

² ARIMA: autoregressive integrated moving average: trung bình động tích hợp tự hồi quy.

Các mô hình ARIMA không theo mùa, tức không có xu thế hay hiệu ứng theo mùa, thường được biểu thị là ARIMA (p, d, q) trong đó các tham số p, d , và q là các số nguyên dương không âm (i) p là thứ tự, tức số lần trễ thời gian, của mô hình tự hồi quy; (ii) d là mức độ sai khác, tức số lần dữ liệu đã bị trừ đi các giá trị trong quá khứ; (iii) q là bậc của mô hình trung bình động. Các mô hình ARIMA theo mùa thường được kí hiệu là ARIMA (p, d, q) (P, D, Q) _{m} , trong đó m là số chu kỳ trong mỗi mùa, và chữ hoa P, D, Q đề cập đến sự tự hồi quy, và thuật ngữ về trung bình động cho phần theo mùa của mô hình ARIMA.

Khi hai trong ba hạng thức là 0, mô hình có thể được gọi dựa trên tham số khác không, đưa về AR, I hay MA. Chẳng hạn ARIMA (1, 0, 0) là AR (1), ARIMA (0, 1, 0) là I (1), và ARIMA (0, 0, 1) là MA (1). Các mô hình ARIMA có thể được ước lượng theo cách tiếp cận Box - Jenkins¹.

Việc kết hợp sự khác biệt của chuỗi thời gian không tĩnh với mô hình ARMA cho ra họ các mô hình, áp dụng trong phạm vi rộng của các tình huống. Sự phát triển của mô hình mở rộng của mô hình này chủ yếu là do Box và Jenkins, nên các mô hình ARIMA còn được gọi là mô hình Box - Jenkins.

Bước đầu tiên trong thủ tục Box - Jenkins là lấy hiệu số trên chuỗi thời gian cho đến khi nó tĩnh, do đó đảm bảo loại trừ các xu thế và các thành phần theo mùa. Trong nhiều trường hợp, chỉ cần đến một hoặc hai lần lấy hiệu số là đủ. Chuỗi lấy hiệu số này sẽ ngắn hơn chuỗi ban đầu theo c bước thời gian; trong đó c là khoảng lấy hiệu số. Khi đó, mô hình ARMA khít với chuỗi thời gian tạo ra.

Bởi vì các mô hình ARIMA có ba tham số, có nhiều biến thể để khớp mô hình. Tuy nhiên, người ta xác định những tham số này theo một số nguyên tắc cơ bản: (i) mô hình nên càng đơn giản

¹ Tiếp cận Box - Jenkins do G. Box và G. Jenkins đề xuất năm 1970, ứng dụng cho mô hình ARMA, hay ARIMA.

càng tốt, tức là có chứa càng ít hạng thức càng tốt, tức p và q nên nhỏ; (ii) phù hợp với dữ liệu lịch sử tốt nhất có thể, tức giảm sai khác bình phương giữa giá trị ước lượng tại bất kỳ khoảng thời gian trước và giá trị thực; (iii) tự tương quan thành phần đo được tại các trễ 1, 2, 3... cần cung cấp chỉ số bậc của thành phần AR, tức giá trị được chọn cho q ; (iv) hình dạng đồ thị của hàm tự tương quan ACF có thể gợi ý kiểu mô hình ARIMA cần thiết. Bảng dưới đây của NIST¹ hướng dẫn việc giải thích mẫu ACF khi lựa chọn mô hình.

Bảng 4.12. Dùng dạng hàm ACF để chọn loại mô hình ARIMA

Hình dạng ACF	Mô hình chỉ định
Hàm mũ, phân rã về 0	Mô hình tự hồi quy. Sử dụng đường tự tương quan bộ phận để xác định bậc của mô hình tự hồi quy.
Luân chuyển âm và dương, dần về 0	Mô hình tự hồi quy. Dùng đường tự tương quan để xác định bậc.
Một hay nhiều đỉnh nhọn, phần còn lại là 0	Mô hình trung bình động. Xác định bậc tại nơi đường chạm 0.
Suy giảm, bắt đầu sau một chút trễ.	Hỗn hợp mô hình trung bình động và mô hình tự hồi quy.
Tất cả là 0, hoặc gần 0.	Dữ liệu là ngẫu nhiên.
Giá trị cao tại khoảng cố định.	Có hạng thức tự hồi quy theo mùa.
Không suy giảm về 0.	Chuỗi không tính.

Các mô hình chuẩn ARIMA thường được mô tả bằng bộ ba (p, d, q) như đã nói ở trên. Chúng xác định cấu trúc của mô hình theo (i) bậc mô hình AR; (ii) lấy hiệu; (iii) bậc MA. Một số dạng thông dụng:

¹ NIST: National Institute of Standards and Technology: Viện Tiêu chuẩn và Kĩ thuật Quốc gia Hoa Kỳ.

- ARIMA (0, 1, 0), hay I(1), cho $x_t = x_{t-1} + z_t$, tức bước đi ngẫu nhiên;
- ARIMA (0, 1, 0) với hằng số, cho $x_t = c + x_{t-1} + z_t$, tức bước đi ngẫu nhiên có lệch;
- ARIMA (0, 0, 0) là mô hình ồn trắng¹;
- ARIMA (0, 1, 2) là mô hình Holt;
- ARIMA (0, 1, 1) với hằng số là mô hình làm mịn theo hàm mũ;
- ARIMA (0, 2, 2) cho $x_t = 2x_{t-1} - x_{t-2} + (\alpha + \beta - 2)z_t - 1 + (1 - \alpha)z_{t-2} + z_t$, tương đương với phương pháp tuyến tính Holt, với sai số phụ, hay làm mịn theo hai lần số mũ.

Cũng có thể có thông số tương tự cho mùa trong dữ liệu, mặc dù các mô hình như vậy sẽ phức tạp hơn khi cần khớp mô hình và giải thích. Bộ ba (P, D, Q) thường được sử dụng để xác định các thành phần mô hình như vậy. Hình sau cho thấy SPSS chọn thủ công các phần tử cấu trúc không theo mùa và theo mùa.

Thông thường các mô hình ARIMA được sử dụng để dự đoán, đặc biệt trong lĩnh vực mô hình kinh tế vĩ mô và mô hình kinh tế vi mô. Tuy nhiên, chúng có thể được áp dụng trong nhiều phạm vi, với các biến số dự đoán bổ sung, để cải thiện độ tin cậy của các dự đoán. Điều này rất quan trọng bởi vì toàn bộ cấu trúc của các mô hình ARMA đã thảo luận ở trên phụ thuộc vào các giá trị trước và các sự kiện ngẫu nhiên độc lập theo thời gian chứ không dựa trên các yếu tố giải thích, tác nhân. Do đó các mô hình ARIMA sẽ chỉ phản ánh và mở rộng các mô hình trong quá khứ, nó cần được sửa đổi trong dự đoán, thông qua các yếu tố như môi trường kinh tế vĩ mô, thay đổi công nghệ, hoặc thay đổi nguồn tài nguyên môi trường.

¹ Ổn trắng: là một dạng tiếng ồn được tạo ra bằng cách kết hợp nhiều loại âm thanh với tần số khác biệt lại với nhau.

4. Một số kĩ thuật phân tích dữ liệu

Time Series Modeler: ARIMA Criteria

Model: Outliers

ARIMA Orders

Structure

	Nonseasonal	Seasonal
Autoregressive (p)	0	0
Difference (d)	0	0
Moving Average (q)	0	0

Current periodicity: 7

Transformation

☒ None

☐ Square root

☐ Natural log

☒ Include constant in model

Continue Cancel Help

4.30. Mô hình ARIMA, chế độ chuyên gia, trong phần mềm :

ình ARIMA trong dự đoán: mô hình ARIMA có thể được xây dựng từ hai tầng mô hình. Trước tiên là làm tính chuỗi $(1-L)^d x_t$. Sau đó làm tính với nghĩa rộng. $(1 - \sum_{i=1}^p \phi_i L^i) (1-L)^d x_t$. Lúc này có thể thực hiện quá trình dự đoán y_t , sau đó tính x_t bằng cách nhân dự đoán tự hồi quy.

hoảng dự đoán, tức khoảng tin cậy cho các dự đoán từ mô hình ARIMA dựa trên giả thiết rằng các phần dư là độc lập và có phân phối chuẩn. Nếu không đúng hai giả thiết này, thì các khoảng dự đoán có thể không chính xác. Vì lý do này, các nhà nghiên cứu vẽ ACF và biểu đồ $\hat{\sigma}^2$ còn lại để kiểm tra các giả thiết trước khi đưa ra dự đoán.

biến thể và phân mở rộng: một số biến thể về mô hình ARIMA được sử dụng. Nếu dùng nhiều chuỗi thời gian x_t như vectơ, và mô hình VARIMA là thích hợp.

khi nghi ngờ có một hiệu ứng mùa trong mô hình; nói chung tốt hơn là sử dụng mô hình SARIMA, tức ARIMA theo mùa, hơn là tăng bậc của phần AR hoặc MA trong mô hình. Nếu nghi chuỗi thời gian có biểu hiện của phụ thuộc trong dài giá trị lớn, thì tham số d có thể lấy giá trị không nguyên trong mô hình trung bình động tích hợp tự hồi quy từng phần, tức mô hình FARIMA hoặc ARFIMA¹.

Triển khai phần mềm: đã có các gói phần mềm khác nhau áp dụng phương pháp luận như tối ưu hóa tham số Box - Jenkins, cho phép xác định tham số cho mô hình ARIMA:

- EViews², phần mềm trong môi trường Windows. Nó dùng để phân tích kinh tế lượng trên chuỗi thời gian. EViews đã mở rộng khả năng ARIMA và SARIMA;
- Mathematica³, phần mềm tính toán sử dụng trong các lĩnh vực khoa học, kỹ thuật, và toán học cũng như các lĩnh vực tin học khác; có hàm về quá trình ARIMA;
- MATLAB⁴, phần mềm cung cấp môi trường tính toán số và lập trình, với công cụ kinh tế lượng, có mô hình ARIMA và hồi quy với sai số ARIMA;
- IBM SPSS bao gồm mô hình ARIMA trong các gói thống kê và mô hình. Tính năng Expert Modeler mặc định đánh giá tự hồi quy theo mùa hoặc không theo mùa, tức p , tích hợp, tức d , và trung bình động, tức q , theo bảy mô hình làm mịn theo số mũ.
- Dịch vụ phân tích của SQL Server của Microsoft bao gồm ARIMA như là một thuật toán khai phá dữ liệu.

¹ AFRIMA, FARIMA: autoregressive fractionally integrated moving average model: mô hình trung bình động tích hợp tự hồi quy từng phần.

² EViews năm 1994.

³ Mathematica do Stephen Wolfram tạo ra năm 1988.

⁴ MATLAB: MATrix LABoratory, được Cleve Moler phát minh vào cuối thập niên 1970.

4.4.5. Phân tích phổ

Phần trước đã cho biết phân tích các chuỗi thời gian trong miền thời gian. Một chuỗi thời gian có hoặc được cho là có chu kỳ phức tạp cũng có thể được phân tích trong miền tần số, với các thủ tục đặc biệt để kiểm tra các mẫu tần số trong chuỗi. Hình thức phân tích này, đôi khi được gọi là *phân tích điều hòa*, *phân tích phổ* hoặc phân tích Fourier, tương tự như các phương pháp *tự tương quan* và *tự hồi quy*, nhưng sử dụng các hàm lượng giác để mô hình hóa dữ liệu.

Phân tích phổ có một mối quan hệ trực tiếp với các phương pháp *tự hiệp phương sai* và các phương pháp *tự tương quan*. Tuy nhiên, trong bối cảnh của chuỗi thời gian gặp phải trong nhiều ứng dụng, trừ những vấn đề liên quan đến phân tích bức xạ điện từ EM, kĩ thuật truyền thông và một số quá trình phức tạp trong môi trường tự nhiên. Lí do là phương pháp này đơn giản và hiệu quả hơn, so với công cụ của phương pháp miền thời gian.

Trong mô hình tự hồi quy, có thể mô hình hóa giá trị x tại thời điểm t như tổng có trọng số của các giá trị trước, cộng thêm lỗi hiện thời, tức $x_t = \alpha_1 x_{t-1} + \dots + \alpha_p x_{t-p} + z_t$, hay $x_t = \sum_{i=1}^p \alpha_i x_{t-i} + z_t$. Nếu nghi ngờ dữ liệu chuỗi thời gian chứa dải chu kì khác nhau p_i , ứng với thời lượng khác nhau, thì mỗi p_i sẽ xảy ra với tần số $\omega_i = 1 / p_i$ chu kì thời gian. Do đó, nếu có các chu kỳ 3, 6 và 12 tháng trong chuỗi dữ liệu theo tháng, chúng ứng với các tần số $1/3$, $1/6$ và $1/12$. Các hàm tự nhiên để mô hình hóa tính chu kì là các biểu thức lượng giác, đặc biệt là các hàm sin và cos. Nếu không điều chỉnh, các này này dao động tron quanh trục ngang, tức trục thời gian trong chuỗi thời gian t , với dải giá trị $[-1, 1]$. Vì $\sin(0) = 0$, hàm này vượt qua trục t tại 0 và $0 \pm \pi$. Hàm cos tương tự, nhưng có giá trị 1 tại $t = 0$. Để có tỉ lệ và chuyển một hàm sin hoặc cos để phù hợp với các mẫu quan sát trong chuỗi thời gian, cần có thêm tham số (i) r : tham số tỉ lệ, làm tăng cường độ, tức biên độ, của hàm; (ii) ω : tham số tần số, làm thay đổi số các chu kỳ mà mà thực hiện trong bội lần π ; (iii) θ : tham số để trượt mô hình của chu kỳ sang trái hoặc phải dọc theo trục, để chúng vượt qua

trục tại các điểm mong muốn, tức thay đổi pha của hàm. Các thông số bổ sung này có thể kết hợp vào một mô hình chuỗi thời gian đơn giản ban đầu trong miền tần số: $x_t = \sum_{i=1}^k r_i \cos(\omega_i t + \theta_i) + z_t$. Biểu thức này cũng có thể được biểu hiện bằng các hàm sin. Do $\cos(\omega_i t + \theta_i) = \cos(\omega_i t) \cos(\theta_i) - \sin(\omega_i t) \sin(\theta_i)$ nên $x_t = \sum_i \cos(\omega_i t) \cos(\theta_i) + \sum_i \sin(\omega_i t) \sin(\theta_i)$, và thay a, b cho phân tử biên độ, pha, và thêm hạng thức hằng số về trung bình, thu được: $x_t = \mu + \sum_i a_i \cos(\omega_i t) + \sum_i b_i \sin(\omega_i t) + z_t$.

Mô hình này đại diện cho x tại thời điểm t bằng hai biểu thức, mỗi biểu thức đại diện cho tổng có trọng số của một số hữu hạn k tần số. Mô hình này có thể được mở rộng đến một số vô hạn của các tần số cho phép $k \rightarrow \infty$, với thể hiện biểu hiện kết quả là tích phân, thay cho tổng. Theo dạng này, không tính hạng thức về lỗi ngẫu nhiên, mô hình được gọi là *thể hiện phổ*¹ của quá trình. Dạng thông thường của thể hiện là: $x_t = \int_0^\pi \cos(\omega t) du(\omega) + \int_0^\pi \sin(\omega t) dv(\omega)$; trong đó u và v là các quá trình liên tục trong đoạn $[0, \pi]$. Nó cho biết tự hiệp phương sai tại trễ k của chuỗi thời gian rời rạc $\gamma(k)$ gần với hạng thức đầu tiên trong thể hiện phổ. Vậy tự hiệp phương sai, và phương sai là: $\gamma(k) = \int_0^\pi \cos(\omega k) f(\omega) d\omega$ và $\gamma(0) = \int_0^\pi f(\omega) d\omega = \sigma_x^2$; trong đó $f(\omega)$ là hàm mật độ phổ. Nhiều gói phần mềm cho phép phân tích phổ với tùy chọn vẽ về các hàm mật độ phổ như là giá trị góc ω , thay đổi từ 0 đến π . Cụ thể, $f(\omega)d\omega$ thể hiện sự tham gia vào phương sai tổng thể, đã được tạo ra bởi các tần số trong dải $[\omega, \omega + d\omega]$, do đó một đỉnh phổ trong hàm mật độ phổ tương ứng với sự tham gia của phương sai trong miền đó. Phương sai tổng thể là tích phân của hàm mật độ phổ, cũng là tự hiệp phương sai tại độ trễ 0. Tỉ số hệ số tự tương quan tại độ trễ k và độ trễ 0 là hệ số tự tương quan. Hai thể hiện, theo miền thời gian và theo miền tần số, do đó xem là tương đương.

Trong thực tế tổng quát hơn, thể hiện chuẩn được sử dụng trong phân tích phổ, gọi là chuỗi Fourier hữu hạn hoặc thể hiện

¹ Lag: tự hậu, trễ.

điều hòa: $x_t = a_0 + \sum_{p=1}^{\frac{N}{2}-1} (a_p \cos(\frac{2\pi p t}{N}) + b_p \sin(\frac{2\pi p t}{N})) + a_{\frac{N}{2}} \cos(\pi t)$,

$t = 1..N$; trong đó $a_0 = \bar{x}$, $a_{\frac{N}{2}} = \sum_t \frac{(-1)^t x_t}{N}$, $a_p = \frac{2}{N} \sum_t x_t \cos(\frac{2\pi p t}{N})$,

và $b_p = \frac{2}{N} \sum_t x_t \sin(\frac{2\pi p t}{N})$. Công thức này có N tham số và được sử

dụng để phù hợp với N điểm dữ liệu, do đó, có một phù hợp chính xác và không có hạng thức lỗi. Các hệ số cho mỗi giá trị tần số có thể thu được như là nghiệm bình phương tối thiểu cho tập gốc các biểu thức, tức những hạng thức đó bao gồm hạng thức lỗi z_t . Biểu thức $r_p^2 = a_p^2 + b_p^2$ là biên độ bình phương của tần số, hay điều hòa, thứ p . Với một chuỗi thời gian có chiều dài N , sự tham gia của điều

hòa thứ p trong tổng bình phương là: $I(\omega_p) = \frac{N r_p^2}{4\pi}$, $p \neq N/2$.

Biểu thức này có thể được vẽ ra theo ω , để có đồ thị cho thấy sự đóng góp của các tần số khác nhau thay đổi khi thay đổi tần số. Đồ thị này được gọi là *biểu đồ chu kì*¹, dù đây có nhầm lẫn, vì nó cho thấy tần số chứ không phải cho khoảng thời gian. Các gói phần mềm phân tích phổ cung cấp sơ đồ chu kì làm công cụ chuẩn để phân tích phổ. Giá trị cao trong khoảng thời gian có nghĩa là những tần số đó có thay đổi nhiều hơn so với các tần số khác.

4.5. Kết luận

Khai phá dữ liệu xử lý nhiều loại dữ liệu. Chương này đề cập loại dữ liệu (i) số, cả số liên tục lẫn rời rạc; (ii) dữ liệu phạm trù; (iii) dữ liệu dạng đồ thị và (iv) chuỗi thời gian. Trong phạm vi tài liệu này, dữ liệu nhân sẽ đề cập sau.

Việc xét các loại dữ liệu là do nhu cầu ứng dụng làm việc với các loại dữ liệu đó. Dữ liệu chuỗi thời gian là loại dữ liệu đặc biệt; nó cần được xem xét kỹ hơn, kèm với các bộ phần mềm xử lý chuỗi thời gian.

¹ Periodogram: biểu đồ chu kì.

Mô hình phân tích và khai phá dữ liệu

5.1. Các mô hình và khuôn mẫu

5.1.1. Giới thiệu về mô hình và khuôn mẫu

Đối với khai phá dữ liệu người ta phân biệt (i) mô hình và (ii) khuôn mẫu¹. Các lớp chính của các mô hình và khuôn mẫu dùng trong khai phá dữ liệu.

Định nghĩa: Mô hình là mô tả mức cao, tổng quát về bộ dữ liệu. Nó cho biết hướng của mẫu lớn. Nó cho biết cách tóm tắt dữ liệu theo cách ngắn gọn và phù hợp, hay có thể được suy diễn, cho phép người ta khẳng định về dân cư mà mô hình được rút ra hay các giá trị tương lai.

Dưới đây sẽ xét các dạng cơ bản như (i) các mô hình hồi quy; (ii) các mô hình hỗn hợp; và (iii) các mô hình Markov.

Định nghĩa: Khuôn mẫu là khía cạnh cục bộ của dữ liệu, có thể giữ vài bản ghi hay một ít biến, hay cả hai. Khuôn mẫu có thể là khía cạnh "cấu trúc" trong không gian các biến p chiều như một kiểu trong hàm mật độ hay điểm cong trên đường hồi quy.

Thường quan tâm đến khuôn mẫu do chúng thể hiện sự khởi đầu của dữ liệu, chẳng hạn (i) cặp các biến với độ tương quan cao; (ii) tập các chi tiết có giá trị cao bất thường trong các biến; hay (iii) nhóm các bản ghi luôn tập trung vào một số biến.

¹ Model: mô hình; Pattern: khuôn mẫu.

Người ta muốn xác định các phần tử để lập luận, mô tả. Cần xác định các phần tử trong cơ sở dữ liệu đang dùng mà mang các tính chất khác thường. Việc nén dữ liệu có thể minh họa khuôn mẫu và mô hình.

- Khi nén ảnh người ta dùng *mô hình*, để ảnh nén đầu kia được giải nén. Theo mô hình về ảnh, người ta thu được ảnh toàn thể, tuy không như ảnh ban đầu;
- Theo tiếp cận *khuôn mẫu*, người ta chỉ xét tập trung vào một phần của ảnh. Đương nhiên đầu kia không có được ảnh ban đầu, nhưng chú trọng vào một khía cạnh đã được lấy mẫu thì người ta có nhiều thông tin hơn tiếp cận mô hình.

5.1.2. Cơ bản của mô hình hóa

Một mô hình là thể hiện trừu tượng của quá trình thực. Chẳng hạn $Y = 3X + 2$ cho biết Y liên quan X , là thể hiện của mô hình $Y = aX + b$. Để có yếu tố ngẫu nhiên trong quan hệ, có thể dùng $Y = aX + b + e$; biến e là biến ngẫu nhiên, $\theta = \{a, b\}$ là tham số.

Trước hết người ta chọn dạng mô hình, rồi xác định tham số.

Xây dựng mô hình trong khai phá dữ liệu là quá trình hướng dữ liệu. Tuy nhiên tất cả các mô hình đều chưa chắc đúng, trừ một vài cái. Do dựa trên dữ liệu, việc phát hiện các mô hình dự đoán cao không có nghĩa do quan hệ nhân quả. Chẳng hạn phân tích dữ liệu về khách hàng mua rượu chất lượng cao không có nghĩa họ mua nhiều quần áo. Tuy nhiên khi có nhiều dữ liệu thì mô hình chính xác hơn. Do không có quan hệ nhân quả, việc xử lý biến này sẽ không ảnh hưởng đến biến khác.

5.1.3. Các cấu trúc mô hình dự đoán

Trong mô hình dự đoán, một trong các biến được mô tả như hàm của các biến khác. Điều này cho phép biến trả lời Y có giá trị theo giá trị của các biến khác X_i , tức các biến giải thích, hay dự

đoán. Mô hình cho phép dự đoán $\hat{y} = f(x_1, \dots, x_p, \theta)$; trong đó $\hat{y}$ là dự đoán của mô hình, θ là tham số cấu trúc mô hình.

- Khi Y nhận giá trị *định lượng*, nhiệm vụ ước lượng việc khớp X_p chiều sang Y được gọi là hồi quy;
- Khi Y nhận giá trị *hạng mục*, tức *phạm trù*, việc học để khớp X sang Y được gọi là học phân lớp, hay phân lớp có giám sát;
- Cả hai nhiệm vụ này được xem như bài toán xấp xỉ hàm, để khớp X_p chiều với Y.

5.1.3.1. Các mô hình hồi quy với cấu trúc tuyến tính

Người ta xét $\hat{Y} = a_0 + \sum_{j=1}^p a_j X_j$, trong đó $\theta = \{a_0, \dots, a_p\}$ là tham số. Mô hình dùng để dự đoán khớp, nhưng không ngầm quan hệ nhân quả. Người ta dùng $\hat{Y}$ thay vì Y bởi lẽ trong mô hình, $\hat{Y}$ được kiến thiết từ các dữ liệu. $\hat{Y}$ được ước lượng từ X , chứ không là giá trị quan sát thực sự. Có thể vẫn giữ bản chất mô hình, mà thêm hàm trong tổng thay cho X_j , $\hat{Y} = a_0 + \sum_{j=1}^p a_j f_j(X_j)$. Các hàm f_j là liên tục, nhưng có thể phi tuyến.

Người ta có thể tổng quát hóa cấu trúc mô hình tuyến tính để đa thức trong các X theo thuật ngữ tích chéo, cho phép tương tác. Việc cho phép mô hình có thuật ngữ thứ hạng cao hơn và tương tác giữa các thành phần của X , về nguyên tắc người ta có thể ước lượng khía cạnh phức tạp hơn nhờ mô hình tuyến tính đơn giản.

Việc tổng quát hóa khối đa thức dẫn đến khái niệm về độ phức tạp của mô hình. Mô hình phức tạp chứa mô hình đơn giản, như lồng nhau. Một mô hình phức tạp luôn có thể khớp với dữ liệu quan sát như bất kì mô hình đơn giản hơn, do nó bao mô hình đơn giản.

Mô hình hồi quy tuyến tính được coi như tìm kiếm giá trị kì vọng của phân phối Y theo giá trị dự đoán X, kí hiệu là $E[Y|X]$.

Mô hình dự đoán cho phép dự đoán theo tham số của phân phối có điều kiện của Y , trong đó có tham số.

5.1.3.2. Các cấu trúc mô hình hồi quy cục bộ từng phần

Định nghĩa: Mô hình cục bộ từng phần là mô hình chung xét trong phạm vi hạn chế của không gian X , tức Y được xét với X cục bộ.

Về hình học, cấu trúc mô hình gồm tập các siêu phẳng p chiều, mỗi cấu trúc phủ một vùng của đầu vào X , không giao với vùng khác. Ý tưởng này được tổng quát hóa cho nhiều biến dự đoán. Mỗi đoạn cục bộ, tức siêu phẳng chứ không đơn giản như đường, điểm, có thể giao nhau. Các cấu trúc cây cũng được dùng cho cấu trúc này.

Loại mô hình tuyến tính cục bộ cũng giúp giải thích việc xây dựng mô hình phức tạp dần dần, dựa trên các mô hình đơn giản. Trong khai phá dữ liệu, việc xây dựng mô hình tiến hành theo quá trình lặp, cũng như phân tích dữ liệu.

5.1.3.3. Các mô hình cục bộ không tham số, dựa vào bộ nhớ

Một trong cách cục bộ hóa các biến trong mô hình là xét một số mảng dữ liệu được lưu trữ. Cũng có thể xem cách này như khuôn mẫu cục bộ, với cấu trúc dữ liệu cô lập. Các mô hình *đường gồm các điểm chốt* và mô hình *cây* đã thay thế các *điểm dữ liệu* bằng hàm ước lượng qua các điểm dữ liệu láng giềng.

Có thể mở rộng việc này đối với các đối tượng của bộ dữ liệu, nhưng sử dụng trọng số về mức độ tương tự so với đối tượng mới. Trong loại mô hình này, người ta cũng áp dụng ngữ cảnh ước lượng mô tả và mật độ. Một vấn đề trong loại mô hình này là xác định *trọng số*. Hàm trọng số cho phép ước lượng mịn, liên tục khi thay đổi chậm khi giảm độ tương tự, ước lượng *lớn chồm* khi thay đổi nhanh. Hàm trọng số được chia thành hai nhóm (i) dạng hàm chính xác của nó; (ii) bằng thông. Giả sử hàm trơn là $K((x - z) / h)$, xác định ước lượng phân phối của ước lượng điểm z mới theo điểm dữ liệu x . Kích thước của phân phối tùy thuộc vào dạng của K

và kích thước bằng thông h . Giá trị h càng lớn cho hàm ước lượng càng mịn; h càng nhỏ ước lượng càng thô.

Các *phương pháp nhân*¹ liên quan đến các phương pháp người láng giềng gần nhất; cả hai lớp phương pháp đều được phát triển. Trong khi các phương pháp nhân xác định cấp độ mịn theo nghĩa hàm nhân và bằng thông, phương pháp người láng giềng gần nhất cho phép xác định bằng thông theo nghĩa số các láng giềng gần nhất. Các cấu trúc cục bộ như các mô hình nhân thường được mô tả là không tham số, do mô hình này hướng dữ liệu không tham số theo nghĩa thường, trừ bằng thông h . Các kỹ thuật làm mịn hướng dữ liệu, như phương pháp nhân, hữu hiệu đối với diễn giải dữ liệu, ít ra cho dữ liệu hai chiều.

5.1.3.4. Các thành phần ngẫu nhiên của các cấu trúc mô hình

Với dữ liệu thực, không thể tìm thấy mối quan hệ hoàn hảo giữa các giá trị dự đoán X và biến trả lời Y . Phân phối của giá trị Y với mỗi giá trị của X thể hiện khía cạnh biến thiên, mà không giảm được, dù bằng mô hình lí tưởng hơn.

Khi tìm Y , các biến thiên của Y có thể được giải thích theo X , người ta dùng khái niệm *các thành phần hệ thống, giải thích được*. Khi không giải thích được, có thuật ngữ về *không hệ thống, không giải thích được*, tức ngẫu nhiên.

Các thành phần mô hình là ngẫu nhiên do (i) sai số độ đo; (ii) tập biến X không bao gồm hết các biến cần thiết. Tuy nhiên, người ta thường làm việc với tập các biến X được thực tế thể hiện. Thành phần ngẫu nhiên trong mô hình là quan trọng khi dùng nó để chọn hàm đích, để ước lượng, để so sánh các mô hình.

5.1.3.5. Các mô hình dự đoán để phân lớp

Khi Y lấy giá trị hạng mục, tức phạm trù, bài toán *dự đoán* chuyển về *phân lớp*. Đích của bài toán là gán đối tượng mới vào

¹ Kernel: nhân.

lớp đúng của chúng, dựa trên các giá trị X quan sát. Trong phân lớp, người ta quan tâm đến mô hình hóa các biên giữa các lớp. Như với bài toán hồi quy, có thể dùng các giả thiết về tham số đơn giản đối với dạng hàm của đường biên. Ở đây xác định biên cho siêu phẳng p chiều.

Mô hình phân lớp cũng được chia theo đơn giản, phức tạp, và tổng hợp mô hình phức tạp dần dần từ mô hình đơn giản. Có nhiều kĩ thuật phân lớp khác nhau để phân lớp biên: (i) láng giềng gần nhất là uyển chuyển, đối với vùng không giao đối với mỗi lớp; (ii) mô hình đơn giản như siêu phẳng đơn giản.

Theo quan điểm mô hình hóa thực hành, hiểu biết trước về đáng điệu của đường biên không thực sự có ý nghĩa đối với bài toán phân lớp, so với ý nghĩa của tri thức về Y liên quan đến X trong bài toán hồi quy. Tuy nhiên, các dạng hàm dùng cho các mô hình phân biệt là hoàn toàn tương tự như trong mô hình hồi quy.

5.1.3.6. Lựa chọn mô hình phù hợp

Mô hình cục bộ từng phần có độ phức tạp tùy theo số k các miền dữ liệu. Khi k lớn, bài toán trở nên phức tạp. Khi tăng năng lực của mô hình, người ta nghĩ đến mức độ phù hợp với dữ liệu. Tuy nhiên, khi bộ dữ liệu huấn luyện tăng, có thể mô hình không tốt như trước. Mặt khác, khi người ta chỉ quan tâm đến việc đơn giản việc mô hình hóa, kết quả thu được lại quá đơn giản.

Việc chọn giữa đơn giản và phức tạp được xem xét. Theo quan điểm hướng dữ liệu, tức trong khai phá dữ liệu, người ta có thể xác định hàm đích cho phép ước lượng mô hình có khớp với dữ liệu mới, nhưng không trên bộ dữ liệu huấn luyện. Người ta dùng (i) tiếp cận thông thường là *kết hợp* phần phù hợp với dữ liệu huấn luyện và dữ liệu thực; (ii) tiếp cận *chia dữ liệu* thành các tập con và huấn luyện các mô hình trên các bộ dữ liệu con.

5.1.4. Các mô hình về phân phối xác suất và các hàm mật độ

Trong mô hình dự đoán, người ta xét Y tính theo các giá trị X . Liên quan đến các dữ liệu, không cần chú trọng đến Y , mà chỉ xét

mối quan hệ trong tổng thể các X , người ta chú trọng đến *tính diễn tả* của dữ liệu. Khi đó việc suy luận về cấu trúc mô hình là quan trọng.

5.1.4.1. Các khái niệm chung

Trong mô hình hồi quy, có các hàm liên tục, trơn; nhưng với khái niệm khác, có các hàm thay đổi đáng điều đột ngột. Về dạng hàm cho mô hình, người ta đề cập việc tổ hợp các hàm đơn giản. Có hai lớp mô hình:

1. *Các mô hình tham số*: các mô hình tham số có dạng hàm giả định cụ thể. Hàm này được đặc trưng qua tham số vị trí và tham số vô hướng, cho biết giá trị biến thiên. Mô hình tham số có ưu điểm là đơn giản; nhưng có nhược điểm là nhập nhằng, vì dữ liệu không theo dạng hàm giả định;
2. *Các mô hình không tham số*: các mô hình không tham số có phân phối và mật độ dựa trên dữ liệu, và ít nhiều cũng có giả thiết trước về dạng hàm.

5.1.4.2. Hỗn hợp các mô hình tham số

Mật độ hỗn hợp đối với x được xác định
$$p(x) = \sum_{k=1}^K p_k(x|\theta_k)\pi_k.$$

Mô hình này phân rã mật độ toàn thể, hay phân phối, đối với x thành tổ hợp tuyến tính có trọng số của K thành phần, hay lớp. Mỗi mật độ thành phần $p_k(x|\theta_k)$ bao gồm điển hình mô hình tham số đơn giản, như phân phối chuẩn, với tham số θ_k . Kí hiệu π_k thể hiện xác suất để điểm dữ liệu chọn ngẫu nhiên được tạo cho thành phần k , $\sum \pi_k = 1$.

Người ta dùng tính toán độ lệch sai phân; không hạn chế chỉ sử dụng thành phần chuẩn, dù có khuynh hướng áp dụng chuẩn cho toàn thể dân cư. Nguyên tắc chung cho mô hình hỗn hợp là ứng dụng rộng rãi và ý tưởng chung là xây dựng mô hình xác suất.

Theo nghĩa diễn giải, (i) hoặc các mô hình hỗn hợp được dùng đơn giản như hộp đen để được dạng mô hình mềm dẻo; hoặc (ii)

các thành phần hỗn hợp cá nhân với diễn giải rõ. Một diễn giải của mô hình hỗn hợp, đặc biệt trong việc phân cụm, là các thành phần được tạo bởi (i) biến ẩn nhận giá trị K ; (ii) vị trí và hình dạng trong không gian p chiều của các thành phần biết trước, thông qua dữ liệu. Vậy nên các mô hình hỗn hợp chia sẻ với định hướng và phương pháp liên quan, về ý tưởng chung. Ý tưởng này nhằm đặt giả thuyết về vấn đề đơn giản hay cấu trúc ẩn được tạo nên nhờ các dữ liệu quan sát.

5.1.4.3. Phân phối kết nối đối với dữ liệu phạm trù không thứ tự

Đối với dữ liệu phạm trù, người ta dùng hàm phân phối nối trong tích chéo của tất cả các giá trị có thể của p biến riêng. Chẳng hạn A là biến có giá trị $\{a_1, a_2, a_3\}$, B có $\{b_1, b_2\}$, phân phối nối của A và B có sáu giá trị có thể. Giả sử các giá trị thực sự là phạm trù, không có hướng và thứ tự.

- Với giá trị p nhỏ, ít biến, người ta dễ thể hiện các giá trị phân phối ở dạng bảng ngẫu nhiên các ô; mỗi ô chứa một giá trị nối. Tuy vậy việc này không thực tế;
- Với các biến phạm trù không có thứ tự, không có hàm phân phối xác suất trơn. Chẳng hạn các biến có m giá trị có thể, cần $m^p - 1$ giá trị xác suất độc lập để xác định toàn bộ mô hình.

5.1.4.4. Yếu tố chung và tính độc lập trong các chiều lớn

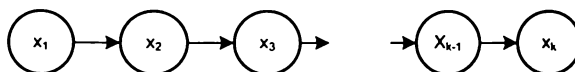
Vấn đề số chiều tác động nhiều đến phân phối và mật độ. Khi số chiều tăng nhanh, khó có thể kiến thiết hoàn hảo các mô hình đơn giản do độ phức tạp tăng theo hàm mũ so với chiều. Xác định yếu tố chung của hàm mật độ rồi lấy thừa số chung cho phép thu được các thành phần thuận tiện đối với việc kiến thiết mô hình. Chẳng hạn khi các biến độc lập, có thể viết hàm mật độ

nối $p(x) = p(x_1, \dots, x_p) = \prod_{k=1}^p p_k(x_k)$, trong đó $x = (x_1, \dots, x_p)$, p_k là

hàm mật độ một chiều. Việc lấy thừa số chung không khó, với giả thiết các biến độc lập và không xấp xỉ với nhiều vấn đề thực.

Tuy nhiên hàm xác suất nối $p(x)$ được viết tổng quát là $p(x) = p_1(x_1) \prod_{k=2}^p p_k(x_k | x_1, \dots, x_{k-1})$ có các thành phần là xác suất có điều kiện.

Đôi khi người ta dùng kinh nghiệm để khớp các mô hình khác nhau với dữ liệu để đơn giản hóa việc lấy thừa số chung. Cũng có thể việc đơn giản này được thực hiện với cấu trúc dữ liệu; chẳng hạn khi các biến thể hiện cùng độ đo tuần tự, như thời gian. Lúc này có thể dùng xích Markov là phù hợp, trong đó phần tử sau chỉ phụ thuộc vào một phần tử trước $p(x_k | x_1, \dots, x_{k-1}) = p(x_k, x_{k-1})$.

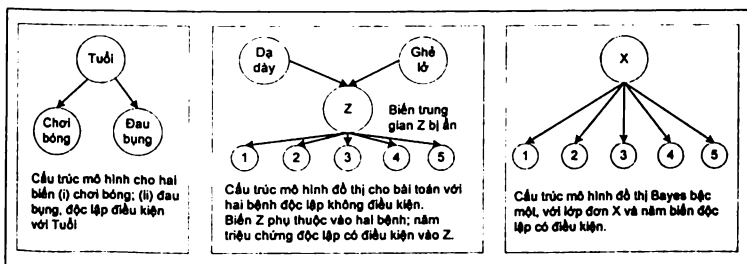


Hình 5.1. Cấu trúc mô hình đồ họa ứng với giả thiết Markov bậc một

Theo ngữ nghĩa của phân lớp và phân cụm, giả thiết các biến là độc lập có điều kiện với các biến khác, tức $p(x | y) = \prod_{j=1}^p p_j(x_j | y)$,

trong đó y là giá trị lớp cụ thể, theo phạm trù. Đây là dạng mô hình Bayes đơn giản. Công thức này có thể dùng với Y là biến không quan sát được, được dùng để đơn giản hóa mô hình $p(x)$,

tức người ta có hỗn hợp vô hạn có dạng $p(x) = \sum_{k=1}^K (\prod_{j=1}^p p_j(x_j | y = k)) \cdot p(y = k)$, trong đó Y lấy K giá trị, và mỗi thành phần $p(x | y = k)$ được mô hình hóa với giả thiết độc lập có điều kiện nêu trên.



Hình 5.2. Các cấu trúc mô hình

5.1.5. Khắc phục số chiều

Thông thường dữ liệu tăng theo hàm mũ khi số chiều tăng. Bài toán tìm mô hình hay khuôn mẫu đều cần dữ liệu, nhưng không thể xử lý lượng lớn các dữ liệu.

Hai chiến lược cơ sở khắc phục khó khăn về số chiều (i) dùng tập con các biến để kiến thiết mô hình, tức chỉ dùng p' biến thay vì p biến, $p' \ll p$; (ii) chuyển p biến thành tập p' biến, $p' \ll p$.

5.1.5.1. Lựa chọn các biến cho dữ liệu nhiều chiều

Lựa chọn các biến là chiến lược tổng quát, nhạy cảm trong bài toán dữ liệu nhiều chiều. Không để tất cả p biến tham gia vào xác định đầu ra Y . Người ta có thể dùng khái niệm về độc lập để hình dung định lượng. Nếu $p(y|x_1) = p(y)$, hay $p(y|x_1, x_2) = p(y|x_2)$ thì biến X_1 không ảnh hưởng đến quyết định Y .

Nếu Y là biến phạm trù, như trong phân lớp, có thể đo thông tin tương hỗ trung bình giữa Y và X' $I(Y; X') = \sum_{i,j} p(y_i, x_j)$.

$\log \frac{p(y_i, x_j)}{p(y_i) \cdot p(x_j)}$ để có ước lượng về sự phụ thuộc giữa X và Y ;

trong đó X' là biến phạm trù, như lượng hóa giá trị thực X .

Trong thực tế, các phương pháp lựa chọn tập con có xu hướng dùng phương pháp may rủi để tìm cấu trúc mô hình tốt. Nhiều thuật toán dựa trên may rủi đơn giản như lựa chọn tham lam, như thêm, bớt một biến đồng thời.

5.1.5.2. Chuyển hóa dữ liệu nhiều chiều

Khắc phục dữ liệu nhiều chiều có thể nhờ chuyển hóa các biến dự đoán. Ý tưởng đơn giản là tìm tập p' biến, gọi là $Z_1, \dots, Z_{p'}$, $p' \ll p$, Z là hàm của các biến X ban đầu. Việc thay các biến quan sát bằng tập các biến ít hơn được dùng trong nhiều ngành của phân tích dữ liệu. Các biến Z được xem như các hàm cơ sở, các nhân tố, các biến rõ, các thành phần chính... phụ thuộc vào các đích và phương pháp sử dụng chúng.

Xét vài mô hình qua hai thí dụ (i) hồi quy theo hướng chiếu; (ii) phân tích thành phần chính.

1. *Hồi quy theo hướng chiếu*: Mô hình có cấu trúc $\hat{y} = \sum_{j=1}^p w_j h_j(\alpha_j^T x)$, trong đó $\alpha_j^T x$ là chiếu của vectơ x trên vectơ trọng số α_j ; cả hai vectơ đều là p chiều được tạo do phép nhân trong, h_j là hàm phi tuyến của chiếu vô hướng này, w_j là các trọng số vô hướng đối với các hàm phi tuyến. Các thủ tục xác định w_j , dạng của h_j , và hướng chiếu α_j có thể phức tạp và tùy theo thuật toán;
2. *Phân tích thành phần chính*: Phân tích thành phần chính PCA là kĩ thuật cổ điển, cho phép thay thế p biến dự đoán ban đầu bằng tập p biến $(Z_1, \dots, Z_p)$, là tổ hợp tuyến tính của các biến ban đầu. Phân tích thành phần chính là trường hợp đặc biệt của hồi quy theo hướng chiếu, nhưng được xem là kĩ thuật giảm dữ liệu. Người ta thấy (i) nó rút ra lần lượt các thay đổi nhất của dữ liệu trong không gian X , nên hi vọng các biến đầu tiên này, ít hơn p biến, mang nhiều thông tin trong dữ liệu; (ii) các thành phần rút ra là trực giao, vậy dễ giải thích. Kĩ thuật PCA được sử dụng nhiều trong phân tích dữ liệu và là công cụ giảm chiều hữu hiệu.

5.1.6. Các mô hình về cấu trúc dữ liệu

Trong nhiều trường hợp, hoặc các cá nhân, hay các biến, hoặc cả cá nhân và các biến có quan hệ trước. Hơn nữa có thể chúng mang quan hệ phức tạp. Cấu trúc dữ liệu một phần chịu ảnh hưởng của phép đo đã xét trước đây. Người ta đã giả thiết các số đo độc lập; tuy nhiên khó đảm bảo tính độc lập. Tùy theo ứng dụng mà có giả thiết cụ thể về dữ liệu. Bản thân thông tin về sự phụ thuộc dữ liệu cũng cần được mô hình hóa.

Về số chiều của dữ liệu, đã có các tiếp cận để giảm khó khăn về nhiều chiều. Mỗi quan hệ và khía cạnh riêng của dữ liệu cần được lấy ra.

Các tiếp cận tổng quát, về chuyển các số đo tuần tự sang dạng vector không tuần tự được xem là đủ cho một số nhiệm vụ khai phá dữ liệu. Tuy nhiên nó làm mất thông tin của các số đo ban đầu, ngay cả thời gian và thứ tự thể hiện thông tin.

Bài toán đơn giản cần cấu trúc dữ liệu đơn giản, chẳng hạn bài toán hai chiều, Y phụ thuộc X . Nhưng cũng có bài toán cần cấu trúc dữ liệu phức tạp hơn như dữ liệu không gian, thời gian, hay lưới các dữ liệu.

5.1.7. Các cấu trúc khuôn mẫu

Khuôn mẫu đặc trưng cho khía cạnh cục bộ của dữ liệu. Khuôn mẫu có thể được xem như vị từ, cho giá trị đúng khi các đối tượng hay một phần của đối tượng trong dữ liệu có mẫu xuất hiện, ngược lại cho giá trị sai.

Khuôn mẫu cần được xác định theo (i) cú pháp của khuôn mẫu, tức ngôn ngữ đặc tả cách xác định khuôn mẫu; (ii) ngữ nghĩa, tức diễn giải về dữ liệu và ứng dụng của dữ liệu. Phần này đề cập khuôn mẫu theo hai dạng khác nhau của dữ liệu giá trị rời rạc (i) dạng ma trận; (ii) dạng sâu.

5.1.7.1. Khuôn mẫu trong các ma trận dữ liệu

Tiếp cận tổng quát xác định khuôn mẫu là (i) bắt đầu từ khuôn mẫu sơ khởi; (ii) kết hợp theo liên kết logic. Tuy nhiên có thể xây dựng khuôn mẫu đặc biệt cho ứng dụng riêng.

Về ma trận dữ liệu: giả sử có p biến $X_1, \dots, X_p$. Lấy $x = (x_1, \dots, x_p)$ là vector p chiều trong bộ dữ liệu $x(i)$, $i = 1..n$. Toàn thể dữ liệu kí hiệu là $D = \{x(1), \dots, x(n)\}$. Và $X_k(i)$ là giá trị số đo thứ k trên cá thể i .

Định nghĩa: Khuôn mẫu đối với các biến $X_1, \dots, X_p$ là tập con của các quan sát có thể trên các biến.

Một ngôn ngữ tổng quát thể hiện khuôn mẫu được xây dựng dựa trên các khuôn mẫu sơ khởi. Có các điều kiện đơn giản trên giá trị các biến. Chẳng hạn c là một giá trị của X_k , thì $X_k = c$ là một

khuôn mẫu sơ khởi. Nếu giá trị của X_k được sắp thứ tự, người ta có $X_k < c$ cũng là điều kiện sơ khởi. Cho tập các khuôn mẫu sơ khởi, người ta xác định khuôn mẫu phức tạp nhờ AND, OR.

Định nghĩa: Lớp khuôn mẫu là tập các khuôn mẫu hợp lệ. Lớp khuôn mẫu C được xác định bằng tập các khuôn mẫu sơ khởi, và cách thức hợp lệ cho phép kết hợp các khuôn mẫu này.

Thí dụ các biến X_i có giá trị trên $\{0, 1\}$. Lớp C là $(X_{j1} = 1) \wedge (X_{j2} = 1) \wedge \dots \wedge (X_{jk} = 1)$. Các khuôn mẫu trong lớp này xuất hiện thường xuyên trong bộ dữ liệu D sẽ được gọi là các tập thường xuyên, của các biến, vì mỗi khuôn mẫu này được xác định duy nhất bằng tập con các biến: khuôn mẫu này có thể được viết như $\{X_{j1}, X_{j2}, \dots, X_{jk}\}$.

Định nghĩa: Tần suất $fr(\rho)$ của khuôn mẫu ρ là số tương đối các xuất hiện trong bộ dữ liệu mà ρ đúng.

Những thuộc tính quan trọng của khuôn mẫu là (i) tần suất $fr(\rho)$; (ii) tính đơn giản của ngữ nghĩa; (iii) tính dễ hiểu; (iv) tính mới hay mức tác động của khuôn mẫu.

Định nghĩa: Phát hiện khuôn mẫu là, với lớp các khuôn mẫu, cần tìm ra tất cả các khuôn mẫu thỏa mãn dữ liệu, vẫn đáp ứng yêu cầu của bộ dữ liệu.

Chẳng hạn tìm tập khuôn mẫu thường xuyên với tần suất ít nhất 0.1 với xác suất xuất hiện trong khuôn mẫu.

Nói chung, định nghĩa nhiệm vụ phát hiện khuôn mẫu có điều kiện về thông tin, tính mới, mức độ dễ hiểu. Do vậy khi phát hiện khuôn mẫu, cần cân đối giữa tính thể hiện của khuôn mẫu, mức phù hợp, độ phức tạp của thuật toán tìm khuôn mẫu. Cho lớp khuôn mẫu C. Người ta xác định các luật. Các luật này có thể nhận các giá trị chắc chắn.

Định nghĩa: Luật là biểu thức $\rho \rightarrow \varphi$, trong đó ρ , φ là các khuôn mẫu trong lớp C, với ngữ nghĩa: nếu ρ đúng đối với một đối tượng, thì φ cũng đúng.

Mức độ đúng của luật được xác định qua xác suất có điều kiện $p(\rho|\varphi)$. Có thể ước lượng xác suất này qua tần suất

$\hat{p}(\varphi | \rho) = \frac{fr(\varphi \wedge \rho)}{fr(\rho)}$. Trợ giúp $fr(\rho \rightarrow \varphi)$ của luật $\rho \rightarrow \varphi$ được xác định

hoặc $p(\rho)$, tức tỉ lệ các đối tượng có luật áp dụng, hoặc $fr(\rho \wedge \varphi)$, tức tỉ lệ các đối tượng có các vế trái, phải của luật đều đúng. Chẳng hạn có luật $\{A_1, \dots, A_k\} \rightarrow \{B_1, \dots, B_h\}$; trong đó A_i, B_i là biến nhị phân. Viết lại luật ở dạng $(A_1 = 1), \dots, (A_k = 1) \rightarrow (B_1 = 1), \dots, (B_h = 1)$. Luật này gọi là luật kết hợp, ứng dụng trong khai phá dữ liệu.

Các khuôn mẫu nêu trên xác định với tập con của bộ dữ liệu. Các công thức xác định chỉ với các biến quan sát đơn giản. Trong trường hợp này, cần có khuôn mẫu trên nhiều quan sát. Việc quan sát tất cả dẫn đến khái niệm phụ thuộc hàm trong mô hình dữ liệu quan hệ. Các khuôn mẫu và điều kiện này chỉ tham chiếu đến các giá trị xuất hiện trong bản ghi đơn trong cơ sở dữ liệu. Đôi khi người ta cũng mô tả khuôn mẫu hướng đến quan sát khác, chẳng hạn tìm người có thu nhập ít nhất.

5.1.7.2. Các khuôn mẫu đối với xâu

Loại dữ liệu khác nhau yêu cầu có khuôn mẫu khác. Xét thí dụ về khuôn mẫu cho các xâu. Về hình thức, xâu trên bộ kí tự S là dãy $a_1, \dots, a_n$ các phần tử của S . Tập S có thể gồm các phần tử lấy giá trị nhị phân, tập mã ASCII. Tập tất cả các xâu xây dựng từ S là S^* . Đối với xâu, không có tập cố định các biến. Nếu muốn mô tả kí pháp xác suất với xâu dữ liệu, người ta coi mỗi phần tử của xâu là một biến ngẫu nhiên.

Dữ liệu có thể là một hay nhiều xâu, điều quan tâm là khuôn mẫu xuất hiện mấy lần trong các xâu. Khuôn mẫu xâu đơn giản nhất là xâu con.

Định nghĩa: Khuôn mẫu $b_1, \dots, b_k$ xuất hiện trong xâu $a_1, \dots, a_n$ tại vị trí i nếu $a_{i+j-1} = b_j, \forall j = 1..k$.

Với xâu, người ta quan tâm đến tập lớn hơn các khuôn mẫu. Kí hiệu E là biểu thức chính quy, xác định tập các xâu $L(E)$. Biểu thức E được thể hiện theo các khả năng (i) xâu $s, L(s) = \{s\}$; (ii) nối

kết E_1E_2 , trong đó $L(E_1E_2)$ gồm các xâu nối $L(E_1)$ với $L(E_2)$; (iii) chọn $E_1 \mid E_2$, vậy $L(E_1 \mid E_2) = L(E_1) \cup L(E_2)$; (iv) lặp E^* , vậy $L(E^*)$ gồm các xâu nối kết 0 hay nhiều lần xâu của $L(E)$.

Biểu thức chính quy là dạng khuôn mẫu phù hợp với lớp các xâu. Biểu thức có thể không mô tả được lớp xâu đơn giản, nhưng phù hợp với các trường hợp phức tạp. Biểu thức chính quy tiện để xác định khuôn mẫu trên xâu, nhưng chưa đủ thể hiện các thay đổi của số các sự kiện thời gian. Lớp đơn giản cần đến số lần xuất hiện như các đoạn truyện nhiều kì, có thứ tự cục bộ, tham chiếu đến nhiều biến khác nhau.

5.2. Độ phức tạp của mô hình

5.2.1. Giới thiệu về độ phức tạp của mô hình

Trước một sự kiện cần dự đoán, vấn đề cần giải quyết, người hay nhóm người khác nhau có ý tưởng khác nhau về giải pháp. Người ta có *tập* nhiều mô hình; mỗi cái dùng thuật toán toán học khác nhau để đoán đầu ra. Các thuật toán này xét dữ liệu theo cách khác nhau. Tập các mô hình nêu dưới đây thường có hành vi ít phức tạp như các mô hình đơn, tức chúng kém mềm dẻo trong việc điều chỉnh đối với các thay đổi tùy tiện về dữ liệu huấn luyện, vậy nên có thể tổng quát hóa thành dữ liệu mới chính xác hơn.

Các mô hình tập hợp, như phương pháp *đóng gói*¹, do Breiman đưa ra 1996, *thúc đẩy*², do Freund và Shapire đưa ra năm 1996, hay mô hình lấy trung bình theo *mô hình Bayes*, là phức tạp, được cho là tốt hơn các mô hình thành phần tạo nên của chúng, trên dữ liệu mới. Điều này là nghịch lí, do người ta *ưa dùng các phương án đơn giản hơn*, tức nguyên lí lưỡi dao Occam³. Việc nhận

¹ Bagging: đóng gói.

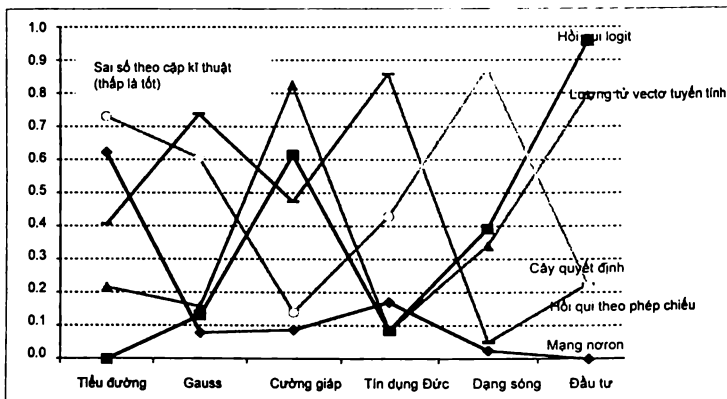
² Boosting: thúc đẩy; là thuật toán siêu tập để học máy, trước tiên giảm thiên lệch vai sai khác trong học giám sát.

³ Occam razor: dao cạo Occam là một nguyên lý do tu sĩ William đề ra, năm 1324, "*nếu không cần thiết đừng tăng thêm thực thể*", tức điều gì có thể được giải

xét tính phức tạp của mô hình (i) không căn cứ vào số các tham số; mà (ii) độ mềm dẻo của quá trình mô hình hóa, tức theo độ tự do tổng quát GDF¹. Do vấn đề chính là đo chức năng mô hình, thay vì dạng của nó, người ta sử dụng được nguyên lí Occam. Chẳng hạn với bài toán cây quyết định hai chiều, tập hợp các cây có độ phức tạp GDF thấp hơn các cây đơn.

5.2.2. Tập mô hình

Có nhiều phương pháp cạnh tranh để rút ra mô hình, và sức mạnh của chúng. Các kết quả phụ thuộc mạnh vào vấn đề. Hình vẽ cho thấy sai số ngoài mẫu² của năm mô hình đối với sáu vấn đề thuộc phạm vi công cộng.



Hình 5.3. Sai số ngoài mẫu tương đối của năm thuật toán, trên sáu vấn đề thuộc phạm vi công cộng

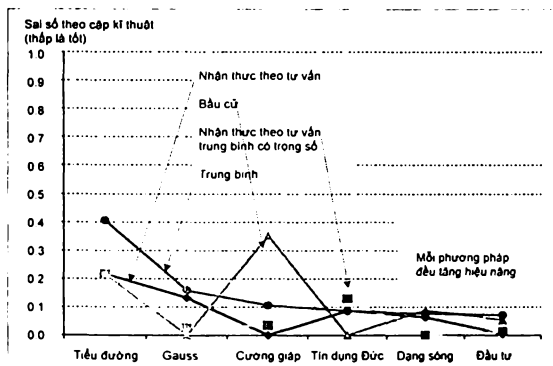
Một phương pháp cải thiện độ chính xác để mạnh hơn việc cắt may thuật toán là: đưa các mô hình vào tập hợp. Hình vẽ sau cho thấy độ chính xác của các thuật toán trong hình vẽ trước, khi

thích bằng ít giả thuyết hơn thì lại được giải thích một cách vô ích bằng nhiều giả định hơn.

¹ GDF: Generalized Degree of Freedom: độ tự do tổng quát.

² Out - of - sample: ngoài mẫu.

kết hợp bốn cách khác nhau, gồm lấy (i) trung bình; (ii) trung bình có trọng số; (iii) bầu chọn; (iv) máy tư vấn.



Hình 5.4. Sai số ngoài mẫu của bốn phương pháp tập hợp theo các vấn đề của hình trước

Tập hợp mô hình được xây dựng theo hai bước (i) kiến thiết các mô hình đa dạng; (ii) kết hợp các ước lượng của chúng. Có thể tạo các mô hình thành phần bằng cách biến đổi các trọng số của các trường hợp, các giá trị dữ liệu, các tham số điều khiển, các tập con của biến, hay các phần của không gian đầu vào. Chẳng hạn *mô hình Bayes* lấy trung bình các ước lượng của các mô hình sẵn có, đánh trọng số theo sự kiện hậu nghiệm; phương pháp *đóng gói*, tức kết hợp các phép thúc đẩy, kích hoạt dữ liệu huấn luyện, thường để xây dựng các cây quyết định, và lấy phiếu bầu cao hay trung bình của các ước lượng; phương pháp *thúc đẩy* lặp lại việc xây dựng mô hình nhờ thay đổi trọng số và sử dụng tổng có trọng số của các ước lượng của dãy các mô hình.

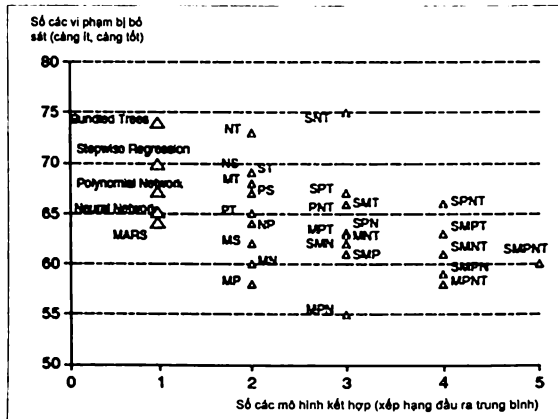
Phương pháp *nhóm việc quản lý dữ liệu* GMDH¹, do Ivakhenko đề xuất 1968, và thế hệ sau là *mạng đa thức*² được xem như các kỹ thuật về tập sớm nhất. Người ta xây dựng nhiều tầng đa thức bậc vừa phải, khớp với hồi quy tuyến tính, mà sự đa dạng xuất hiện từ

¹ GMDH: Group Method of Data Handling; phương pháp nhóm việc quản lý dữ liệu.

² Polynomial Network: mạng đa thức.

các biến khác nhau được dùng tại mỗi nút mạng. Tổ hợp của chúng là phi tuyến, do các đầu ra của nút bên trong là đầu vào cho nút đa thức trong các tầng kế tiếp. Việc kiến thiết mạng dùng lại nhờ thử nghiệm kiểm chứng chéo, dùng GMDH, hay nhờ hàm phạt phức tạp. Phương pháp thông dụng khác, xếp chồng, do Wolpert đề xuất năm 1992, dùng mạng nơron như các thành phần, trong đó tính đa dạng bắt nguồn từ việc sử dụng các tham số điều khiển, chẳng hạn trọng số khởi đầu. Các thành phần kết hợp trong hồi quy tuyến tính, được huấn luyện trên ước lượng *bỏ đi một*¹ từ mạng. Cuối cùng là *hợp nhất mô hình*, do Elder đề xuất năm 1996, đạt được sự đa dạng nhờ lấy trung bình ước lượng của các mô hình được xây dựng từ các thuật toán khác nhau.

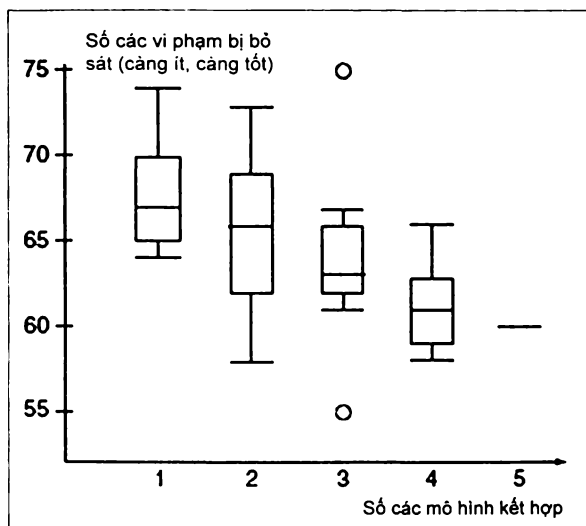
Hình tiếp theo là kết quả ngoài mẫu của việc kết hợp năm dạng mô hình khác nhau trên ứng dụng tính điểm tín dụng. Việc kết hợp xếp thứ tự theo số mô hình tham gia.



Hình 5.5. Lỗi ngoài mẫu về điểm tín dụng, khi trộn một đến năm dạng mô hình khác nhau vào tập

Hình sau cho biết sai số trung bình giảm khi tăng mức kết hợp. Mô hình cuối cùng với năm thành phần tốt hơn bất kỳ mô hình thành phần nào.

¹ Leave-one-out: bỏ đi một.



Hình 5.6. Biểu đồ hộp cho biết sai số trung vị và trung bình, giảm khi mức trộn mô hình tăng lên

5.2.3. Độ phức tạp

Chỉ trích về các *tập* là diễn giải của mô hình sẽ kém đi. Chẳng hạn cây quyết định có các thuộc tính hấp dẫn; hồi quy tuyến tính LR^1 là phương pháp sử dụng rộng rãi... Nhóm các cây vào tập cho phép tăng cạnh tranh tại các thuật toán gay cấn, nhưng mất đi tính diễn giải. Để lượng hóa mất mát này, người ta thấy tập các cây có thể xem như một cây, như từng phần nhỏ trên bề mặt kết quả. Nhưng cây này có nhiều nút hơn các cây thành phần; chẳng hạn đóng gói M cây nhị phân sẽ sinh ra cây 2^M lá.

Tiếp cận trội, do Tibshirani đề xuất năm 1999, cho phép dùng các lợi thế của phương pháp đóng gói, không yêu cầu nhiều mô hình, để giữ lại các diễn giải. Người ta xây dựng các mô hình cạnh tranh từ bộ dữ liệu khởi đầu và chỉ giữ mô hình có ít lỗi nhất trên dữ liệu gốc. Trên bộ dữ liệu mới, người ta thu được kết quả vượt

¹ LR: Linear regression: hồi quy tuyến tính.

trội, chúng tỏ mô hình trội đủ tốt trên cả hai bộ dữ liệu. Tuy nhiên, độ chính xác không tăng như với tập các mô hình.

Chỉ trích khác về tập, ngoài việc chính xác không tăng, là độ phức tạp tăng sẽ dẫn đến quá khít, vậy sẽ không chính xác đối với bộ dữ liệu mới. Thực ra, chưa thấy quá khít đối với tập trong ứng dụng thực tế. Điều này làm người ta nghi ngờ về rằng nguyên lý lưỡi dao Occam đối với tổng quát hóa sẽ gặp trở ngại từ độ phức tạp.

Để thấy tiếp cận tập các mô hình phức tạp ra sao, nên xác định phép đo độ phức tạp. Đối với LR, người ta chỉ tính các hạng thức; việc này không cho phép trong mô hình phi tuyến. Tham số đơn trong mô hình phi tuyến có thể ảnh hưởng ít hơn như trong mô hình tuyến tính, hay lớn hơn vài tiếp cận khác, như ba bậc tác động tự do đối với mỗi tham số trong MARS. Trường hợp đơn giản, *dưới tuyến tính*, có thể xảy ra với mạng nơron khi không huấn luyện đủ các trọng số. Trường hợp *trên tuyến tính* được biết nhiều hơn, chẳng hạn người ta thấy số các bậc tự do liên kết với hồi quy bình phương tối thiểu phi tuyến có thể xem như đáng kể hơn số các tham số yêu cầu khớp.

Số các tham số và bậc tối ưu không phải tất cả những thứ liên quan đến độ phức tạp và tiềm năng về quá khít. Dạng mô hình riêng không lột tả được *mở rộng tìm kiếm đối với cấu trúc*. Chẳng hạn mô hình được giải về khai phá dữ liệu và khám phá tri thức 2001 chỉ dùng ba biến. Nhưng có 140.000 biến ứng viên, với chỉ 2.000 trường hợp. Với tỉ lệ đủ lớn về biến ứng viên duy nhất đối với các trường hợp, các tìm kiếm hạn chế ở việc phát hiện một vài biến có tính thăm dò, cho dù không thấy mối quan hệ thực. Năm 1989 Hjorth có nhận xét rằng *việc đánh giá mô hình được chọn có thể không dựa trên riêng mô hình, mà đòi hỏi thông tin về lớp các mô hình và thủ tục chọn*. Vậy, người ta cần dùng độ đo lựa chọn mô hình, là cái tác động đến việc chọn mô hình.

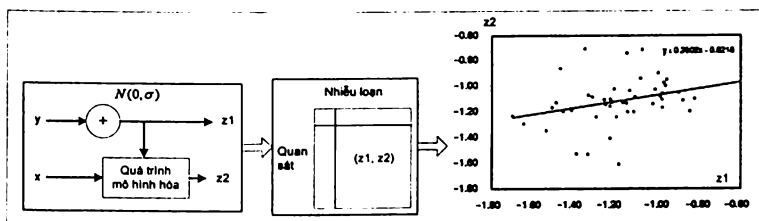
Thực tế cho thấy độ phức tạp được đo không chỉ với mô hình, mà còn toàn bộ *thủ tục* mô hình hóa, tức liên quan trực tiếp đến độ

mềm dẻo của thủ tục. Chẳng hạn tiêu chí lạm phát hiệp phương sai, do Tibshirani và Knight đưa ra năm 1999, khớp với mô hình và tiết kiệm các ước lượng, thì với xáo trộn biến đầu ra, chạy lại thủ tục mô hình hóa, và đo hiệp phương sai giữa các ước lượng mới và cũ. Càng thay đổi lớn, thích nghi với ngẫu nhiên và mềm dẻo, hàm phạt càng phức tạp đòi hỏi huấn luyện lại mô hình này, từ các quá khớp. Đơn giản hơn như tiếp cận GDF, làm nhiều ngẫu nhiên, tức thêm ồn, biến đầu ra, chạy lại thủ tục mô hình hóa, đo thay đổi đối với ước lượng. Tiếp tục, thủ tục mô hình hóa càng thích nghi với ồn thêm vào, các mô hình của nó càng mềm dẻo, tức càng phức tạp.

Bước chính trong việc lặp ngẫu nhiên quanh thủ tục mô hình hóa gọi cho người ta nhớ công cụ phân tích hồi quy, do Faraway đưa ra 1991. Nó cho phép đo, qua việc lấy mẫu lại, độ vững chắc của kết quả từ việc mô hình hóa tự động nhiều bước.

5.2.4. Độ tự do tổng quát

Đối với hồi quy tuyến tính LR, độ tự do K bằng số hạng thức, dù việc này không ngoại suy sang hồi quy phi tuyến. Nhưng có định nghĩa khác $K = \text{vết}(\text{ma trận } \hat{Y}) = \sum \delta Y_hat / \delta Y$; trong đó $\delta Y = Y_e - Y$ và $\delta Y_hat = Y_e_hat - Y_hat$; mà $Y_hat = f(Y, X)$ đối với mô hình $f()$; đầu ra Y , và vectơ đầu vào X ; $Y_e_hat = f(Y_e, X)$; $Y_e = Y + N(0, \sigma_e)$. Đầu ra nhiều loạn ($Y + \text{sai sót}$) sau khi có GDF.



Hình 5.7. Lược đồ của quá trình tính GDF

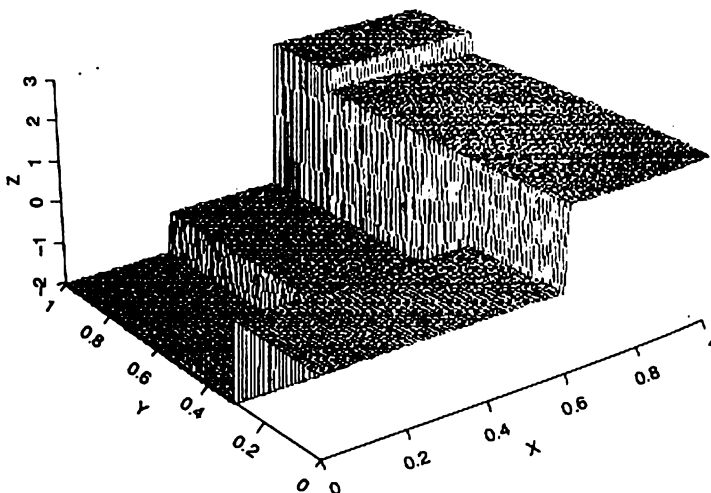
Vậy GDF được xác định là tổng của độ nhạy của mỗi giá trị khớp, Y_hati , để nhiễu loạn trong đầu ra tương ứng, Y_i . Tương tự,

bậc tự do tác động của mô hình đường lười gà được ước lượng theo vết của ma trận chiếu, $S: Y_hat = SY$. Năm 1998, Ye đề nghị tạo bằng nhảy với nhiễu, rồi dùng phương pháp tính GDF *lâm ngang*, như trong hình trên.

Khớp LR với δY_hat_i vào δY_i , khi dùng dòng dữ liệu ứng với trường hợp i , rồi bổ sung độ dốc m_i . Ước lượng này xuất hiện bền vững hơn kết quả của phương pháp *thẳng đứng*, khi lấy trung bình giá trị thu được từ mỗi cột dữ liệu, tức GDF đối với mỗi mô hình hay bộ dữ liệu nhiễu.

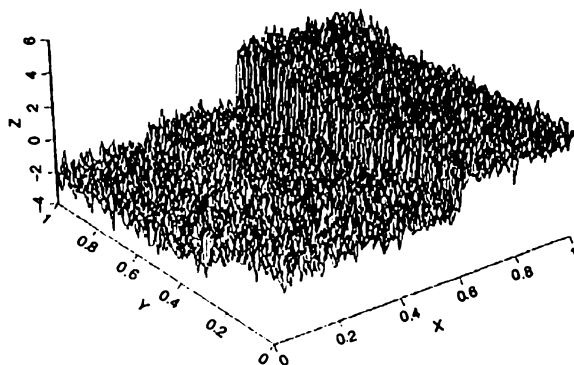
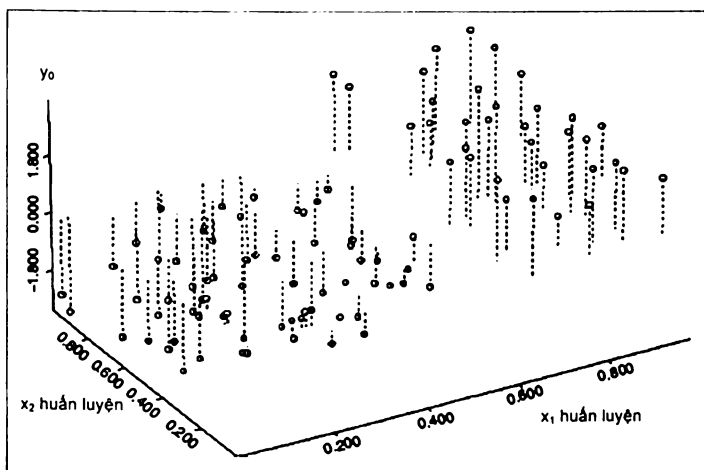
5.2.5. Thí dụ về bề mặt có ồn

Thí dụ này bắt đầu từ thử nghiệm từng mẫu bề mặt cố định hai chiều, để giới thiệu GDF.



Hình 5.8. Bề mặt cây hai chiều dùng trong thực nghiệm

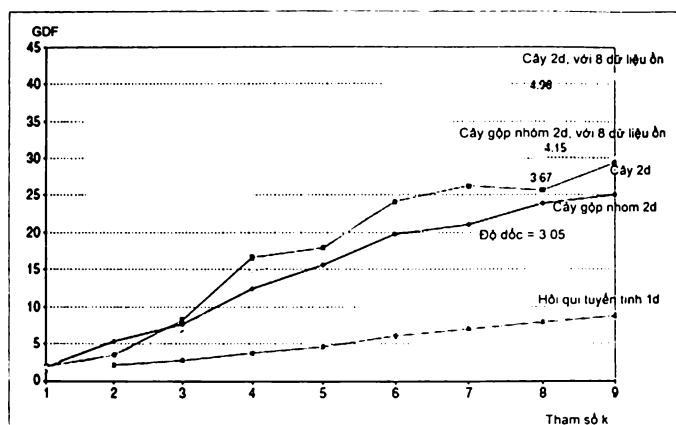
Bề mặt được tạo nhờ cây quyết định, phù hợp với cây, với năm nút lá, tức bốn lần phân chia, có thay đổi cấu trúc nhỏ nhất là 0.5. Hình sau minh họa *bề mặt*, sau khi thêm ồn Gauss $N(0, 0.5)$. Hình tiếp cho biết 100 mẫu ngẫu nhiên của không gian đó.

Hình 5.9. Bề mặt cây, sau khi thêm ồn $N(0, 0.5)$ 

Hình 5.10. Lấy 100 mẫu từ hình trên

Trong hình, dữ liệu ồn trên cây là bộ dữ liệu (X, Y) , dùng cho thực nghiệm. Đối với nhiễu GDF, người ta dùng 50 bản sao, mà mỗi bản thêm vào Y ồn Gauss $N(0, 0.25)$ có nửa độ lệch chuẩn của ồn đã có trong dữ liệu huấn luyện.

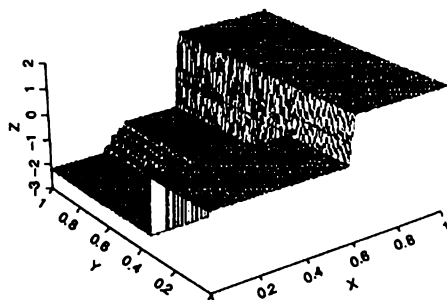
Hình sau cho thấy GDF với dãy K , tức số các tham số, đối với mô hình LR, cây đơn, tập năm cây.



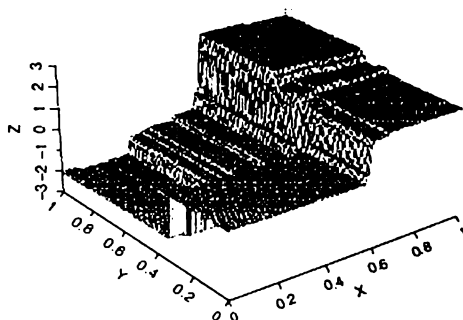
Hình 5.11. Dây GDF đối với năm mô hình, dùng từ một đến chín tham số

Về lý thuyết, GDF cho các mô hình LR khớp chặt chẽ số hạng thức K. Đối với cây quyết định với kích thước K khác nhau, tức số cực đại các ngưỡng phân tách, GDF lớn tới 3.67 lần so với K. Phương pháp đóng gói, tức thúc đẩy việc lấy mẫu bộ dữ liệu và lấy trung bình đầu ra, năm cây với nhau, tỉ lệ phức tạp tăng 3.05. Có lẽ các cây ghép lại với kích thước K đã cho sẽ đơn giản đi năm lần nhờ GDF, so với mỗi thành phần.

Hình sau minh họa hai bề mặt trong dãy các cây đóng gói. Đóng gói năm cây hạn chế tới bốn nút lá, tức ba lần tách, cho bề mặt như hình tiếp theo. Bề mặt với đóng gói phức tạp hơn sẽ mịn hơn.



Hình 5.12. Bề mặt của ghép năm cây, với ba phân tách



Hình 5.13. Bề mặt của ghép năm cây, với bảy phân tách

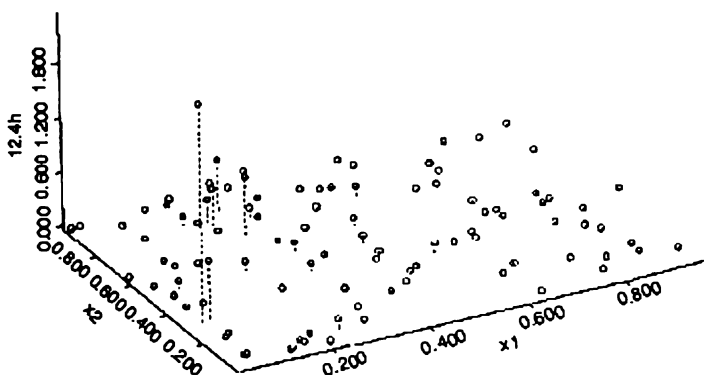
5.2.6. Nhận xét về mô hình

Nhóm các mô hình cạnh tranh thành *tập* luôn cải thiện tính tổng quát, sử dụng các thuật toán khác nhau theo cách hiệu quả, để đạt được tính đa dạng yêu cầu của các thành phần. Các tập *cho thấy* tăng tính phức tạp, do có nhiều tham số hơn so với mô hình thành phần. Việc này vi phạm tính đơn giản trong nguyên lý Occam. Vậy, nếu sử dụng GDF, tức độ đo thực nghiệm về tính *mềm dẻo* và *quá trình* mô hình hóa, để đo tính phức tạp, thì người ta thấy các tập sẽ đơn giản hơn các thành phần. Người ta nhất trí rằng khi đo độ phức tạp phù hợp, thì vẫn giữ được nguyên lý lưới dao Occam.

Với GDF, quá trình mô hình hóa càng khớp với thay đổi tùy tiện để tạo nên đầu ra, nó càng phức tạp. Điều này hợp với lý thuyết tuyến tính, nhưng có thể khác khi so sánh với quá trình mô hình hóa nhiều tầng. Trong thực nghiệm với cây, GDF tăng khi có các biến đầu vào nhiễu, và với năng lực của tham số, khi các cây so với LR. Hi vọng nó cũng tăng với tìm kiếm triệt để, và giảm với mô hình tiên nghiệm, với điều chỉnh tham số, và khi cấu trúc dữ liệu có liên quan rõ với ổn.

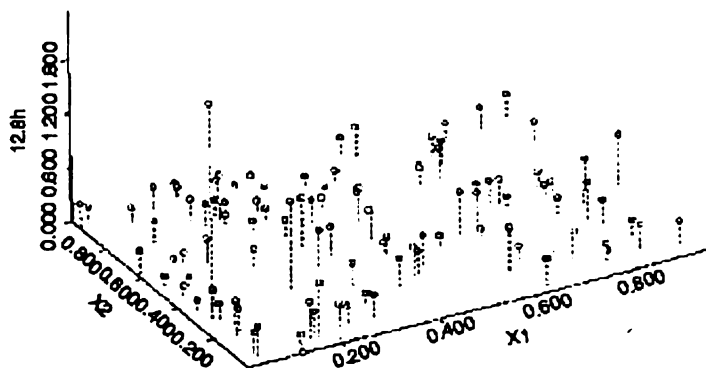
Cuối cùng, tính toán ngang, theo trường hợp, của GDF cho phép xác định sự tham gia vào độ phức tạp của từng trường hợp.

Hình sau cho thấy tác động của hai mô hình cây đơn. Các kết quả cây kém khớp lộ ra chỉ một ít quan sát phức tạp, tức nó dẫn đến các thay đổi trong ước lượng của mô hình khi nhiễu bởi ồn ngẫu nhiên.



Hình 5.14. Tham gia vào độ phức tạp của mỗi mẫu trong ghép năm cây, với ba phân tách

Hình sau cho thấy độ phức tạp khuếch tán đối với kết quả trong trường hợp quá khứ. Thuật toán mô hình hóa sau này sẽ tìm kiếm đệ quy với *gây phức tạp lẽ* và tập trung vào cấu trúc mô hình cục bộ, không dùng mô hình gia tăng trong miền đã ổn định.



Hình 5.15. Gây nên phức tạp của mỗi mẫu trong nhóm ghép của năm cây, với bảy phân tách

5.3. Mô hình phù hợp với mục đích

5.3.1. Giới thiệu về tính phù hợp

Vấn đề tìm kiếm lời giải cho bài toán trong thế giới thực (i) không phải thế giới quá phức tạp; mà (ii) thế giới phức tạp hơn người ta *có thể* nghĩ. Tức giải pháp của con người kém phức tạp như thế giới yêu cầu. Thường đáp án của con người là đơn giản, và một trong các chiến lược là tuân theo suy nghĩ rằng *nhận thức theo cách chung để đơn giản hóa nhiệm vụ*.

Ngày nay, người ta hiểu rằng khả năng công nghệ là ở mọi thứ có thể được, cho phép làm mọi thứ. Lỗi của phê phán này không ở hiểu biết mà con người đã qua trong lịch sử, mà ở giả định rằng người ta *không thể* qua. Không chỉ bản chất phức tạp hơn người ta nghĩ về nó, mà khả năng ứng dụng công nghệ của con người để hiểu nó và giữ nó là lớn hơn nhiều cái người ta tưởng tượng.

Cách người ta xử lý hầu hết vấn đề trong thế giới theo con đường suy nghĩ, theo cách này hay cách khác. Thông thường, người ta là nạn nhân của nhận thức đơn xen về hiện thực. Thay vì dành thời gian để nghiên cứu và hiểu, ít ra về nguyên tắc, tất cả các khía cạnh của hoàn cảnh, người ta nhảy ngay tới *giải pháp để giải vấn đề*, theo con đường rất hẹp, đẩy những ràng buộc, với nhiều giả thiết. Cho dù các giả định được kiểm chứng, người ta bỏ qua tác động của nhiều ảnh hưởng khác, để tạo nên giải pháp trong khung thời gian yêu cầu. Việc tạo giải pháp chấp nhận được trong khung thời gian đã cho là đáng khích lệ, nhưng cách tiếp cận thông thường thì không.

Một trong nhận thức chung trong khai phá dữ liệu là *nhiều hơn tốt hơn*. Điều này diễn tả (i) nhiều là tốt hơn; (ii) cần chọn giữa tính *hiệu quả* hay tính *đủ*.

5.3.2. Bài học từ tự nhiên và kĩ nghệ

Bài học này là *nhiều hơn không nhất thiết là tốt hơn*. *Hiệu quả*¹ được xác định như (i) tối đa đầu ra; (ii) tối thiểu đầu vào. Giải

¹ Hiệu quả xem như *do right* và *right doing*.

pháp, hay nghiệm bài toán, tốt nhất được xác định theo tính hiệu quả: ý tưởng này được tham chiếu đến như *khung hiệu quả*. Khung này giả sử đích của hiệu quả là tối đa đầu ra, trong khi tối thiểu đầu vào.

Trong phân tích thống kê, khung hiệu quả được diễn giải để xác định giải *pháp hiệu quả* như thứ có biến thiên nhỏ tương đối. Một nghiệm, tức một ước lượng, có thể được coi như hiệu quả hơn lời giải khác, nếu hiệu của ma trận hiệp phương sai của lời giải và ma trận phương sai của nghiệm là lớn hơn số dương. Khi tất cả các phần tử ma trận kết quả là dương, nó được gọi là *ma trận nửa xác định dương*. Nghiệm hiệu quả nhất là thứ gần với ma trận nửa xác định dương.

Mô tả thống kê có thể có ích, nếu khung hiệu quả là đúng trong ngữ cảnh của nghiệm. Tuy nhiên, nhiều thí dụ trong thế giới thực cho thấy khác với khung này. Chẳng hạn hậu quả sinh thái xảy ra trong vùng rừng, khi có hiệu quả về thảm cỏ; người ta thấy vai trò của loài cạnh tranh, trong đó có cỏ.

Tiếp cận xác định hiệu quả theo nghĩa *đủ* là lõi của tranh luận năm 2004, theo Falvey, khi xác định nông nghiệp bền vững. Người ta cho rằng cần tránh sai lầm khi tách biệt hiệu quả và tính đủ, mà nên tính đến cả hai. Thay vì để tính đủ phụ thuộc vào tính hiệu quả, người ta đưa ra khái niệm về hiệu quả như là lời giải đủ, không coi nghiệm tối ưu như cái nhất thiết cần đạt được. Theo ngữ cảnh tự nhiên, người ta gọi nó là *khung đủ*¹.

Thậm chí trong kinh doanh, người ta thay khung hiệu quả bằng khung đủ. Khung này trong kinh doanh sẽ phản ánh không chỉ đích kinh doanh của doanh nghiệp, mà còn quá trình kinh doanh để đạt đích.

Nếu theo khung đủ trong thiết kế kho dữ liệu chuyên đề phân tích và phát triển giải pháp khai phá dữ liệu, nó sẽ thay đổi cách người ta tạo các mô hình khai phá dữ liệu. Nó khiến người ta chấp

¹ Efficiency paradigm: khung hiệu quả; Sufficiency paradigm: khung đủ.

nhận lời giải đủ tốt, để xây dựng sự phối hợp và sản phẩm doanh nghiệp; sẽ đạt sản phẩm tối đa *theo điều kiện* hạn chế về đích, để phát triển dài hạn.

Theo khung đủ, nghiệm khai phá dữ liệu tốt nhất sẽ không được xác định đơn lẻ về tối đa đầu ra. Thay vì vậy, người ta xác định theo cách làm việc cùng nhau với quá trình kinh doanh, để cải thiện hành động kết dính trong toàn bộ chuỗi lợi nhuận. Hành động kết dính cho phép doanh nghiệp chủ động và thích nghi với thay đổi, hơn là thụ động và chịu đựng. Tập các đặc trưng này có tính *tổ chức* hơn là *cơ học*.

5.3.3. Ra quyết định trong tổ chức kinh doanh

5.3.3.1. Chấp nhận thay đổi hơn là trốn tránh

Khi bàn về kinh tế sinh học, năm 1991 Rothschild cho rằng tổ chức cơ học sợ thay đổi vì sự kiện xảy ra nhanh hơn quá trình kinh doanh cứng nhắc đáp ứng sự kiện đó. Các doanh nghiệp thiết kế quá trình kinh doanh để có lợi thế về thay đổi đối với tiến triển của thị trường thích hợp. Thông tin có vai trò trong luồng sáng tạo, cho phép tạo nên thông tin tốt. Việc chia sẻ thông tin trong tổ chức kinh doanh sẽ giúp ngăn cản người ta vấp phải lỗi cũ.

5.3.3.2. Trợ giúp quyết định

Định nghĩa: Hệ thống trợ giúp quyết định DSS¹ giúp người quản lý, tức người ra quyết định, thực hiện công tác quản lý.

Thông thường các kết quả khai phá dữ liệu có thể cho phép các hoạt động ra quyết định để thiết kế hành động tại nơi xa của tổ chức. Nhưng các quyết định này có thể khó hay không thể cài đặt được. Chẳng hạn khó tạo lại bản ghi phân tích khách hàng CAR² trong cơ sở dữ liệu tác nghiệp. Lí do là (i) khó truy cập tất cả dữ liệu nguồn đã tạo nên CAR; (ii) không có quá trình kinh doanh

¹ DSS: Decision Support System: hệ thống trợ giúp quyết định.

² CAR: Customer analytic Record: bản ghi phân tích khách hàng.

trong thao tác sản xuất, để chuẩn bị dữ liệu cần thiết, để tạo được CAR. Việc không nối kết giữa mô hình hóa và thao tác nghiệp thể hiện như là ngăn cách trong luồng thông tin. Đó là nguyên nhân nhiều mô hình khai phá dữ liệu xếp trên giá IT, chưa được cài đặt trong sản xuất.

Tổ chức kinh doanh cần có luồng quyết định giữa việc phân tích và sản xuất, cũng như chuyển thông tin nhanh chóng và hiệu quả. Theo Gates, 1999, đó là *hệ thống thần kinh số*.

5.3.3.3. Quan trọng trong tổ chức kinh doanh

Hệ thống quan trọng khác trên đường đến hoạt động kinh doanh là được thể hiện trong quá trình kinh doanh, mà được huấn luyện phù hợp để đưa thông tin quyết định vào hoạt động. Quá trình kinh doanh này tại nơi hành động cần được phát triển *trước* khi nhận được thông tin ra quyết định. Do vậy, *giải pháp khai phá dữ liệu cần có công nghệ đảo ngược, từ điểm hoạt động*. Điều này có nghĩa người ta cần mô hình hóa quá trình phát triển mô hình từ phần cuối trong đơn vị kinh doanh, thay vì từ mặt trước của IT. Các yêu cầu cài đặt cần được thiết kế đối với mỗi bước trên đường của thông tin quyết định, bắt đầu từ điểm hoạt động, và tiến về phía phát triển mô hình. Đầu vào cho mỗi bước cần ghép với việc chuyển hóa thông tin cần thiết, để tạo nên bản chất chính xác của đầu ra của bước đó, đáp ứng đầu ra cho bước tiếp theo.

Thiết kế tổng thể của hệ thống mạng IT và quá trình kinh doanh đối với mỗi bước trên đường là *diễn giải về xây dựng mô hình giải pháp từ trên xuống*, tức theo Plato. Thiết kế chi tiết của quá trình kinh doanh và phân tích tại mỗi bước *thể hiện về xây dựng mô hình giải pháp từ dưới lên*, tức theo Aristotle. Các cam kết và giả thiết cần đảm bảo để qua được vấn đề có thể ngăn cản luồng thông tin. Nếu *xây dựng mô hình giải pháp chỉ từ trên xuống hay từ dưới lên*, người ta sẽ đến nơi mà tự phát hiện rằng chưa đủ biết và chưa đủ hiểu để liên kết các bước trong quá trình. Các kỹ sư hệ thống định lượng các liên kết này theo cách đơn giản và gọi chúng là các hàm

chuyển. Đây là cách các hệ thống phức tạp được mô hình hóa trong thế giới thực.

Định nghĩa: Từ trên xuống dưới và dưới lên là cả hai chiến lược xử lý thông tin và tri thức, được sử dụng trong nhiều lĩnh vực khác nhau bao gồm phân mềm, lý thuyết nhân văn, và khoa học, quản lý và tổ chức.

Tiếp cận từ trên xuống, hay thiết kế theo từng bước, hay phân rã, là sự phá vỡ một hệ thống để có được cái nhìn sâu sắc vào các hệ thống phụ của nó theo kiểu kỹ thuật đào ngược. Tiếp cận từ dưới lên là việc kết hợp các hệ thống để tạo ra các hệ thống phức tạp hơn, do đó làm nổi các hệ thống con của hệ thống ban đầu.

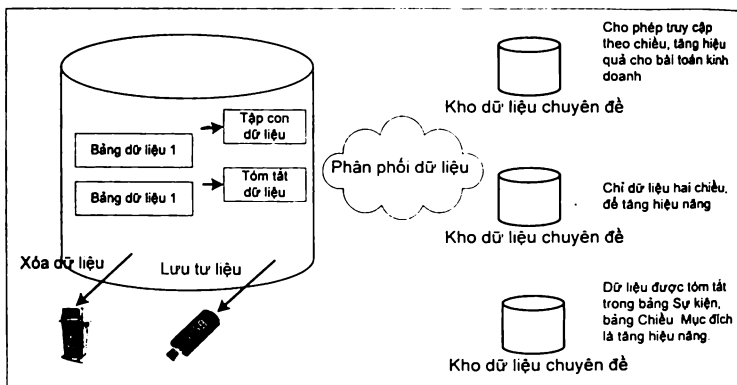
5.3.3.4. Hệ thống phức tạp

Định nghĩa: Hệ thống phức tạp là tổ chức gồm các thành phần liên kết và tương tác với nhau, mà hành vi của chúng từ một phần riêng là không rõ ràng.

Liên quan đến thuộc tính của hệ thống có *thuộc tính nổi trội*. Thuộc tính này là chưa rõ ràng và thậm chí chưa có trong tập các thuộc tính của thành phần hệ thống. Khái niệm về tổ chức kinh doanh có thể xem như ngữ cảnh về hệ thống phức tạp. Để tiện cho các giải pháp khai phá dữ liệu trong tổ chức kinh doanh phù hợp, cách quyết định toàn thể cần được thiết kế cùng với giải pháp trong đầu. Nó có thể được thực hiện hiệu quả trong kho dữ liệu chuyên đề khám phá, phụ thuộc vào kho dữ liệu xí nghiệp. Khái niệm *hệ sinh thái doanh nghiệp* là một phần trong *hệ thống thông tin tích hợp* mà Inmon đưa ra năm 1998. Tiếp cận về kho dữ liệu trợ giúp cho mô hình hóa phân tích.

Kho dữ liệu chuyên đề và các thứ khác được thiết kế để báo cáo, sẽ được thể hiện hoặc dưới dạng logic hoặc vật lý. Kho dữ liệu chuyên đề vật lý được đặt trên hệ thống tách biệt với lược đồ cơ sở dữ liệu dùng cho mục đích chuyên biệt. Kho dữ liệu chuyên đề logic được đặt trên hệ thống kho dữ liệu, và được cài đặt dưới

dạng khung nhìn cơ sở dữ liệu hay nằm trong nhóm các bảng giải chuẩn, mang dữ liệu phù hợp với việc tạo ra CAR.



Hình 5.16. Hệ thống thông tin tích hợp

Thông thường, nơi dễ nhất để bắt đầu là xây dựng kho dữ liệu chuyên đề logic, nhờ thiết kế bảng, với dữ liệu tóm tắt gộp trong mức cá nhân, tổ chức... Các bảng khác mang dữ liệu nhân khẩu có thể thêm vào kho dữ liệu chuyên đề logic, nhờ kết nối các dữ liệu thông qua khóa chung. Cấu trúc logic sẽ phù hợp với nguồn dữ liệu để khai phá dữ liệu.

Nguồn dữ liệu thiết kế riêng không là phần duy nhất trong chuỗi ra quyết định, mà còn có các phần khác:

1. Cấu trúc mạng IT trong công ty là *hệ thống thần kinh số*, nhờ đó mà các kết quả khai phá dữ liệu được liên lạc với các hành động kinh doanh;
2. Quá trình kinh doanh theo điều kiện phù hợp, *chuyển dịch* vụ và chức năng sản xuất của công ty thành lợi nhuận;
3. Hành động đúng được *sinh ra*;
4. Các *kết quả của hành động*, tức đáp ứng của khách hàng, tạo nên ít nhiều giá trị cho công ty;
5. Các kết quả được *phản hồi* về bộ não kinh doanh và đảm bảo cơ sở về việc học cách xây dựng mô hình tốt hơn.

Những đáp ứng học được là đặc tính của tổ chức thích nghi, hơn là cái máy tính, thụ động. Trong hệ thống có tổ chức kinh doanh, các quyết định được thiết kế theo IT đưa đến sự thật. tức là các quyết định không thay đổi về bản chất, khi nó qua nhiều chức năng trong tổ chức kinh doanh, qua thời gian. Lí do chúng không thay đổi là vì chúng được thiết kế đúng, nên chúng khớp với cách ra quyết định trong mỗi quá trình, từ nhận thức trong não đến nơi hành động.

5.3.4. Luật hành động

Định nghĩa: Luật Pareto hay quy luật 80/20, quy luật thiểu số quan trọng và phân phối nhân tố, nói rằng trong nhiều sự kiện, khoảng 80% kết quả là do 20% nguyên nhân gây ra.

Luật, hay nguyên lí, Pareto¹ được dùng để quản lí chất lượng sử dụng năm 1940, lấy khái niệm về ít sống còn, nhiều tâm thường. Việc tạo ra đủ các nghiệm mô hình hóa khai phá dữ liệu có thể theo luật 80/20. Cấu trúc của quá trình mô hình hóa tuân theo luật này: 80% thời gian của dự án mô hình hóa dùng để chuẩn bị dữ liệu, và chỉ khoảng 20% dành cho huấn luyện và thử nghiệm mô hình. Dựa theo luật Pareto, người ta có thể hi vọng đạt đủ nghiệm mô hình trong nhiều trường hợp, với chỉ 20% công sức cho mô hình hóa, để tạo giải pháp với dự đoán tối đa. Có trợ giúp để tin điều này trong khái niệm mô hình hóa nhanh.

5.3.5. Thí dụ về mô hình hóa nhanh

Một trong những tiếp cận sâu sắc về mô hình hóa có nguồn gốc từ môi trường lập trình cực hạn XP². Tiên đề của nó là phân phối nhu cầu khách hàng phần mềm khi cần thiết.

Định nghĩa: Lập trình cực hạn, nhanh, là phương pháp phát triển phần mềm linh hoạt, sử dụng các nhóm làm việc kết hợp gồm (i) người

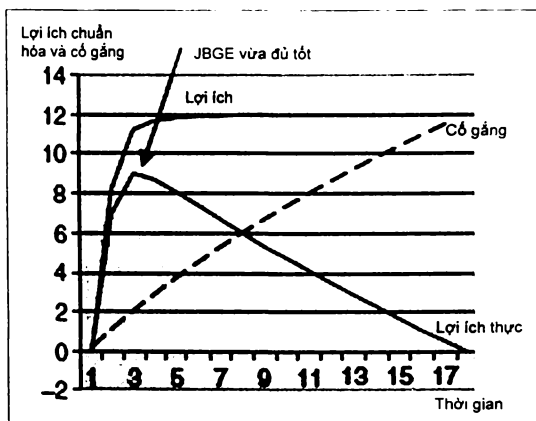
¹ Luật Pareto do Vilfredo Pareto đưa ra năm 1906.

² XP: Extreme Programming: lập trình cực hạn, do Ambler đề xuất năm 2002.

lập trình; (ii) khách hàng; (iii) nhà quản trị, để phát triển phân mềm có chất lượng cao, trong thời gian ngắn.

Một trong chức năng tiện ích quan trọng trong mô hình hóa nhanh là chu trình phản hồi từ cổ đông. Các cổ đông cho quá trình phát triển các điểm quan trọng trong dự án, để kiểm chứng trạng thái hiện thời của tiện ích tiềm năng *về nhận thức*. Ambler để liệt kê sáu đề nghị của mô hình hóa nhanh, liên quan chặt chẽ với việc phát triển mô hình khai phá dữ liệu:

1. *Vừa đủ tốt JBGE¹ là chính sách hiệu quả nhất.* JBGE tương tự như điểm uốn trong đồ thị đáp ứng. Điểm JBGE trong hình sau là vị trí thích hợp nhất. Giả sử cố gắng thêm có thể cho các mô hình JBGE khác, nhưng điểm cắt có lợi về tổng thể vẫn sẽ hạn chế đến mức JBGE này;

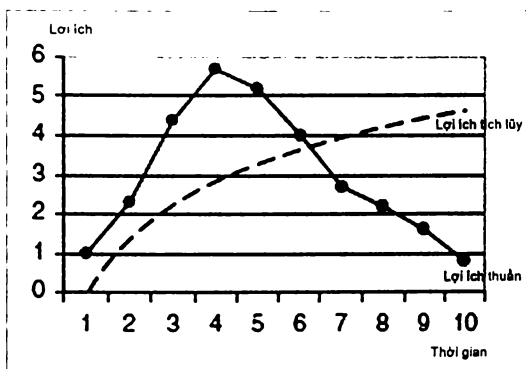


Hình 5.17. Mối quan hệ giữa lợi ích mạng và cố gắng bỏ ra

2. *JBGE không ngầm định chất lượng tốt nhất.* Mức JBGE của việc sinh mô hình có thể không đạt chính xác cao nhất, nhưng đủ để làm việc. Các cổ đông của mô hình này chứng tỏ tốt nhất về tiện ích của mô hình, chứ không phải người lập mô hình;

¹ JBGE Just barely good enough: vừa đủ tốt.

3. *JBGE phụ thuộc vào hoàn cảnh.* Điều đủ tốt cho hoàn cảnh này có thể không tốt cho hoàn cảnh khác;
4. *Mô hình JBGE tiến hóa theo thời gian.* Mô hình ban đầu được làm tinh, hay cập nhật theo thời gian. Khi có nhu cầu, hay khi điều kiện thay đổi, có thể thay đổi mô hình. Các đặc tính của mô hình có thể thích nghi với điều kiện mới. Theo cách này, một mô hình cũng như loài sinh học, đáp ứng với thay đổi điều kiện môi trường, nhờ thay đổi bản chất;
5. *Điểm lợi ích cực đại xuất hiện trước khi người ta nghĩ.* Trong hình sau, điểm lợi ích về mạng ở mức cố gắng 4. Nó có thể xuất hiện không trái trực giác rằng cố gắng bổ sung liên kết với lợi nhuận, nhưng khi đã định được chi phí phụ, mức 4 là tốt nhất;



Hình 5.18. Mối quan hệ giữa giá trị thực hiện, tức lợi ích mạng, và giá trị tích lũy truyền thống.

6. *Giá trị thực hiện của mô hình có thể vượt quá giá trị nhận thức.* Khẳng định này ban đầu có vẻ logic, không ngược với trực giác, nhưng xét tiếp trong ngữ cảnh kinh doanh, người ta thấy sáng rõ hơn. Khái niệm truyền thống của giá trị tăng theo công sức. Nhưng giá trị tiềm năng có thể bị che khuất hay trễ. Giá trị thực hiện của mô hình theo thời gian có thể khác xa mô hình không theo thời gian.

Giá trị tích lũy trong hình thể hiện khung nhìn truyền thống về xác định giá trị theo độ chính xác của mô hình và đặc trưng trong đó. Về bản chất, khi công sức tăng trong phát triển, người ta đạt được chính xác cao hơn và có nhiều đặc trưng thêm vào. Nhưng giá trị tiện ích thực của mô hình này có thể theo đường cong về lợi ích mạng, thay vì đường cong của giá trị tích lũy.

5.4. Phân tích thống kê của dữ liệu bình thường và bất thường

5.4.1. Giới thiệu về phân tích thống kê

Nhiều lĩnh vực phân biệt (i) dữ liệu bình thường, chuẩn; với (ii) dữ liệu bất thường. Chẳng hạn như lĩnh vực chất lượng dữ liệu, khi phân biệt quá trình điều khiển trong, điều khiển ngoài; an toàn dữ liệu phân biệt các hoạt động bất thường từ bên trong hay bên ngoài.

Định nghĩa: Kỹ thuật điều khiển quá trình thống kê là tập các công cụ xác định tốt, để khai phá dữ liệu, phát hiện các mẫu thống kê về dữ liệu bình thường và bất thường.

Điều khiển quá trình thống kê gồm vài phương pháp, cho phép đo, giám sát, và tăng cường chất lượng sản xuất hay quá trình. Các *biểu đồ* điều khiển là công cụ thông dụng nhất để giám sát chất lượng, được W. A. Shewhart đề xuất năm 1920.

Trong quá trình sản xuất luôn có biến động. Biến thiên tự nhiên được nhắc đến là *biến vốn có*², xảy ra khi có cơ hội. Nếu quá trình sản xuất chỉ xuất hiện *cơ hội tạo ra biến thiên*, quá trình này được gọi là hoạt động theo trạng thái điều khiển thống kê. Có vài tình huống mà biến thiên là kết quả của vài thể hiện bất thường hay nhiễu loạn. Khi *nguyên nhân đến gây biến thiên* trong quá trình, người ta nói quá trình ra ngoài điều khiển.

¹ Chart: biểu đồ.

² Inherent: vốn có

5.4.2. Các biểu đồ điều khiển một biến

Các biểu đồ điều khiển một biến dùng để giám sát biến đơn hay thuộc tính quan tâm. Các biểu đồ điều khiển đề cập sau đây gồm (i) $\bar{x}$ và các biểu đồ điều khiển R; (ii) $\bar{x}$ và biểu đồ điều khiển S; (iii) các biểu đồ điều khiển cá nhân; (iv) biểu đồ điều khiển không phù hợp phân số; (v) biểu đồ điều khiển tính không phù hợp.

5.4.2.1. Các biểu đồ điều khiển biến

Biểu đồ điều khiển thông dụng nhất đối với phép đo hay dữ liệu biến thiên là biểu đồ điều khiển của Shewhart, (i) $\bar{x}$ và các biểu đồ điều khiển R; (ii) $\bar{x}$ và biểu đồ điều khiển S. Biểu đồ $\bar{x}$ giám sát trung bình của quá trình; biểu đồ R, S để giám sát đồng thời trung bình của quá trình và biến thiên của quá trình.

Các biểu đồ điều khiển R và $\bar{x}$: khi việc đo theo số thực hiện trên các đặc trưng chất lượng cụ thể, đặc trưng chất lượng được tham chiếu như một *biến*. Người ta dùng $\bar{x}$ và biểu đồ điều khiển R như các kĩ thuật giám sát thống kê đối với dữ liệu biến thiên và thường tham chiếu đến các biểu đồ điều khiển biến.

Xét một biến có phân phối chuẩn, với trung bình μ và độ lệch chuẩn σ . Cả hai giá trị này thường chưa biết, cần được ước lượng. Lấy $x_1, \dots, x_n$ thể hiện mẫu của độ đo kích thước n trên biến quan tâm. Trung bình mẫu này $\bar{x}$ là $\bar{x} = (x_1 + \dots + x_n) / n$; trong đó x có phân phối chuẩn với trung bình μ và sai số chuẩn $\frac{\sigma}{\sqrt{n}}$.

Thực tế các mẫu kích thước n có thể lấy tại các khoảng thời gian thay đổi. Chẳng hạn trong sản xuất bán dẫn, độ mỏng của tấm bìa là đặc trưng quan trọng, cần được giám sát. Người ta chọn ngẫu nhiên một nhóm con các tấm bìa sau mỗi 2 giờ, và đo độ dày. Độ dài trung bình mẫu được tính đối với mỗi mẫu, và trung bình độ dài của tấm bìa được giám sát. Nếu trung bình mẫu dày hay mỏng bất thường, cần báo động cho quá trình.

Giả sử có m nhóm con, có kích thước n , được chọn ngẫu nhiên, để giám sát quá trình cụ thể. Dữ liệu tổng quát như trong bảng kèm theo. Trung bình mẫu và hạng của mẫu thể hiện trong cột ba và bốn. Hạng của mẫu, R_i , là sai khác giữa giá trị cực đại và cực tiểu của mỗi mẫu, và thường dùng để đánh giá biến thiên của mẫu.

Bảng 5.1. Ghi chú tổng quát về dữ liệu nhóm con

Nhóm nhỏ i	Các số đo	$\bar{x}_i$	R_i
1	$x_{11}, x_{21}, \dots, x_{n1}$	$\bar{x}_1$	R_1
2	$x_{12}, x_{22}, \dots, x_{n2}$	$\bar{x}_2$	R_2
3	$x_{13}, x_{23}, \dots, x_{n3}$	$\bar{x}_3$	R_3
$\vdots$	$\vdots \quad \vdots \quad \vdots$	$\vdots$	$\vdots$
m	$x_{1m}, x_{2m}, \dots, x_{nm}$	$\bar{x}_m$	R_m

$$\bar{\bar{x}} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i \quad \bar{R} = \frac{1}{m} \sum_{i=1}^m R_i$$

Thống kê $\bar{\bar{x}}$ xem như trung bình lớn, và là đánh giá chung tốt nhất của trung bình quá trình thực, μ . Hạng trung bình $\bar{R}$ sẽ được sử dụng để đánh giá biến thiên của quá trình và kiến thiết biểu đồ điều khiển cần thiết.

Khi tính được trung bình mẫu, hạng mẫu, trung bình lớn, và hạng trung bình, biểu đồ điều khiển Shewhart được phát triển. Hạn chế điều khiển trên UCL¹, đường giữa CL² và hạn chế điều khiển dưới LCL³ đối với biểu đồ điều khiển $\bar{x}$ là $UCL = \bar{\bar{x}} + A_2 \bar{R}$; $CL = \bar{\bar{x}}$; $LCL = \bar{\bar{x}} - A_2 \bar{R}$; trong đó A_2 là hằng số phụ thuộc vào kích thước mẫu, n . Giá trị A_2 cho trong bảng sau đối với các kích thước mẫu khác nhau.

¹ UCL: upper control limit: hạn chế điều khiển trên.

² CL: centerline: đường giữa.

³ LCL: Lower control limit.

Bảng 5.2. Các hằng số để kiến thiết các biểu đồ điều khiển

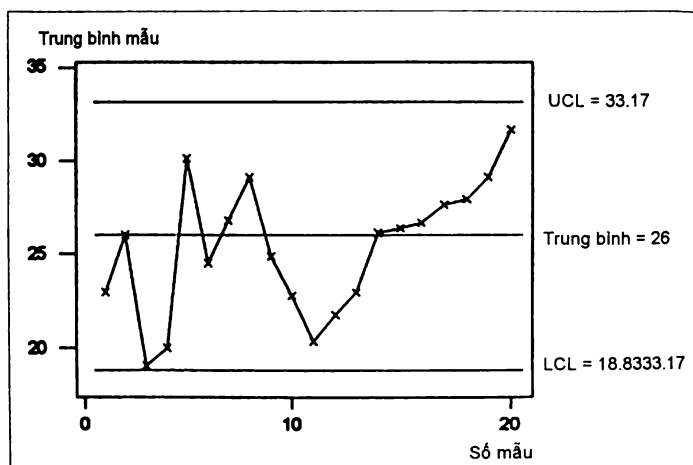
Kích thước mẫu, n	A_2	A_3	B_3	B_5	B_6	d_2	D_2	D_3	D_4
2	1.88	2.659	0	3.267	2.606	1.128	3.686	0	3.267
3	1.023	1.954	0	2.568	2.276	1.693	4.358	0	2.575
4	0.729	1.628	0	2.266	2.088	2.059	4.698	0	2.282
5	0.577	1.427	0	2.089	1.964	2.326	4.918	0	2.115
6	0.483	1.287	0.030	1.97	1.874	2.534	5.078	0	2.004
7	0.419	1.182	0.118	1.882	1.806	2.704	5.204	0.076	1.924
8	0.373	1.182	0.185	1.815	1.751	2.847	5.306	0.136	1.864
9	0.337	1.099	0.239	1.761	1.707	2.970	5.393	0.184	1.816
10	0.308	1.032	0.284	1.716	1.669	3.078	5.469	0.223	1.777

Các UCL, CL và LCL đối với hạng là $UCL = D_4 \bar{R}$; $CL = \bar{R}$; $LCL = D_3 \bar{R}$; trong đó D_3 và D_4 là các hằng số phụ thuộc vào kích thước mẫu, n . Các giá trị D_3 , D_4 cho trong bảng.

Các giá trị tham số rất tiện tính thủ công. Giá trị d_2 thường được tham chiếu như trung bình của độ đo *hạng quan hệ*, xác định là R / σ , phụ thuộc vào kích thước mẫu, n . Công thức của A_2 là $A_2 = \frac{3}{d_2 \sqrt{n}}$, phụ thuộc chỉ vào n , có thể được tính cho các n khác nhau, như trong bảng. Các giá trị tham số D_3 và D_4 được xác định $D_3 = 1 - 3d_3 / d_2$; $D_4 = 1 + 3d_3 / d_2$; trong đó d_3 là hàm của độ lệch chuẩn R . Như A_2 , các giá trị D_3 , D_4 đều là hằng số, phụ thuộc n . Các công thức đối với mỗi tham số cho trong bảng.

Bản thân biểu đồ điều khiển được tạo nên nhờ vạch theo các trung bình của nhóm con, LCL, CL và UCL theo nhóm con, và m . Nếu một hay nhiều điểm trung bình mẫu, hay hạng mẫu, nằm ngoài các hạn chế điều khiển, người ta gọi quá trình là ngoài điều khiển. Các *khuôn mẫu*¹ hay xu thế đã rõ ràng trên các nhóm con cũng yêu tố để xét quá trình ra ngoài vòng điều khiển. Chẳng hạn xét biểu đồ điều khiển trong hình trước. Xu thế đi lên là rõ ràng, bắt đầu từ dữ liệu thứ 11 đến 20. Xu thế này cho biết quá trình không còn sinh các quan sát ngẫu nhiên về trung bình trong vòng điều khiển, và cần nghiên cứu quá trình này về lí do.

¹ Pattern: khuôn mẫu, dễ không trùng với mẫu (sample).



Hình 5.19. Minh họa vẽ biểu đồ điều khiển chuẩn với xu thế đi lên

Về biểu đồ $\bar{x}$ và R , xét bài toán về độ dài bìa bán dẫn như trên. Sau 2 giờ, người ta đo độ dài của 4 tấm bìa. Mười hai nhóm con được chọn, với kết quả đo như trong bảng sau. Kỹ sư sẽ giám sát độ dài trung bình theo dữ liệu.

Giải pháp: Đối với bài toán này $m = 20$, $n = 4$. Để kiến thiết các biểu đồ điều khiển, cần có trung bình mẫu và hạng mẫu của mỗi nhóm con. Để mô tả, trung bình mẫu đối với nhóm 1 là $\bar{x}_1 = (x_1 + \dots + x_4) / 4 = (15.4 + 14.1 + 12.7 + 13.8) / 4 = 14$. Hạng trung bình của nhóm 1 được tính theo $R_1 = \max - \min = 15.4 - 12.7 = 2.7$.

Trung bình mẫu và hạng mẫu đối với tất cả các nhóm cho trong bảng. Để thấy trung bình lớn, người ta tính $\bar{\bar{x}} = (\bar{x}_1 + \dots + \bar{x}_{20}) / 20 = 14.000 + \dots + 14.675 / 20 = 14.96$.

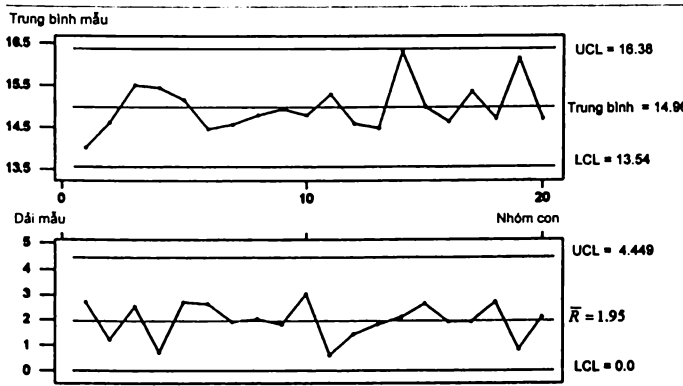
Hạng trung bình là $R = (R_1 + \dots + R_{20}) / 20 = (2.7 + \dots + 2.1) / 20 = 1.95$.

Các giá trị LCL, CL, UCL của biểu đồ điều khiển $\bar{x}$ thu được theo công thức. Trước tiên cần hằng số A_2 , theo bảng dữ liệu, với $n = 4$, người ta lấy $A_2 = 0.729$. Các hạn chế điều khiển và đường trung bình đối với biểu đồ điều khiển $\bar{x}$ là $UCL = \bar{\bar{x}} + A_2 \bar{R} = 14.96 + 0.729 * 1.95 = 16.38$; $CL = 14.96$; $LCL = \bar{\bar{x}} - A_2 \bar{R} = 14.96 - 0.729 * 1.95 = 13.54$.

Bảng 5.3. Độ dài của bản mạch, với trung bình mẫu và khoảng mẫu đối với mỗi nhóm con

Nhóm con i	Độ dài				$\bar{x}_i$	R_i
1	15.4	14.1	12.7	13.8	14.000	2.7
2	15.3	14.2	14.1	14.8	14.600	1.2
3	15.5	15.0	17.0	14.5	15.500	2.5
4	15.7	15.1	15.8	15.1	15.425	0.7
5	14.7	16.8	14.1	15.0	15.150	2.7
6	14.3	15.8	13.2	14.5	14.450	2.6
7	14.3	15.8	14.2	13.9	14.550	1.9
8	14.4	15.3	15.7	13.7	14.775	2.0
9	14.2	16.0	14.4	15.1	14.925	1.8
10	13.2	14.9	16.2	14.8	14.775	3.0
11	15.4	15.6	15.0	15.1	15.275	0.6
12	13.9	14.0	15.3	15.1	14.575	1.4
13	15.2	13.4	15.1	14.2	14.475	1.8
14	17.7	15.6	15.7	16.2	16.300	2.1
15	13.5	14.8	15.5	16.1	14.975	2.6
16	14.3	14.0	15.9	14.3	14.625	1.9
17	14.3	16.2	15.8	15.1	15.350	1.9
18	14.6	15.9	13.2	15.0	14.675	2.7
19	15.7	15.9	16.5	16.4	16.125	0.8
20	15.9	14.9	13.8	14.1	14.675	2.1

Các hạn chế điều khiển và đường trung bình đối với biểu rọc tính theo công thức. Trong bảng dữ liệu có giá trị hằng à D_4 đối với $n = 4$. Ở đây $D_3 = 0$ và $D_4 = 2.282$. Các giá trị $UC = 2.282 * 1.95 = 4.449$; $CL = \bar{R} = 1.95$; $LCL = D_3 \bar{R} = 0 * 1.95$



Hình 5.20. Đồ thị điều khiển R và $\bar{x}$

Các biểu đồ điều khiển kết quả thể hiện trong hình. Quá trình được điều khiển, do không có khuôn mẫu hay xu thế rõ ràng và không có điểm nằm ngoài các hạn chế điều khiển. Các biểu đồ trong hình được kiến thiết nhờ phần mềm Minitab.

Cuối cùng, nếu có điểm ra ngoài các hạn chế điều khiển theo biểu đồ $\bar{x}$ hay biểu đồ R, quá trình này sẽ được xem xét để quyết định phát hiện nguyên nhân hay không. Nếu có điểm được quyết định là bất thường, thì điểm này cần bị loại bỏ; tính lại các hạn chế điều khiển; kiến thiết lại các biểu đồ điều khiển.

Các biểu đồ điều khiển $\bar{x}$ và S: xét tình huống người ta muốn đánh giá nhiều về độ lệch chuẩn mẫu S của quá trình hơn là đánh giá hạng mẫu. Nói chung, độ lệch chuẩn mẫu là đánh giá tốt hơn về độ lệch chuẩn của quá trình thực, khi kích thước mẫu n lớn, tức $n > 10$. Khi n lớn lên, hạng R là kém chính xác và không còn hiệu quả thống kê đối với các kích thước mẫu lớn. Độ lệch chuẩn mẫu cho đánh giá tốt hơn đối với độ lệch chuẩn của quá trình, khi kích thước mẫu không cố định đối với các nhóm con.

Các biểu đồ $\bar{x}$ và S được kiến thiết tương tự như các biểu đồ $\bar{x}$ và R, ngoại trừ độ lệch chuẩn được tính, thay cho việc tính hạng, đối với mỗi nhóm con. Lấy $x_1, \dots, x_n$ thể hiện mẫu kích thước n . Công thức độ lệch chuẩn mẫu của nhóm con i là S_i và độ lệch chuẩn trung bình trên m nhóm con là $\bar{S}$.

$$S_i = \sqrt{\frac{\sum_{j=1}^n (x_{ij} - \bar{x})^2}{n-1}}; \bar{S} = \frac{\sum_{i=1}^m S_i}{m}$$

Các hạn chế điều khiển đối với $\bar{x}$ là $UCL = \bar{\bar{x}} + A_3 \bar{S}$; $CL = \bar{\bar{x}}$; $LCL = \bar{\bar{x}} - A_3 \bar{S}$; trong đó A_3 là hằng số, phụ thuộc vào kích thước mẫu, n , cho trong bảng.

Các hạn chế đối với biểu đồ S là $UCL = B_4 \bar{S}$; $CL = \bar{S}$; $LCL = B_3 \bar{S}$; trong đó B_3 và B_4 là hằng số phụ thuộc vào kích thước mẫu, n , cho trong bảng.

Về biểu đồ $\bar{x}$ và S: để minh họa việc kiến thiết $\bar{x}$ và S, tiếp tục xét thí dụ về độ dày bìa bán dẫn, như trong thí dụ trước. Thay vì

tính hạng R của mỗi nhóm con, cần tính độ lệch chuẩn cho mỗi nhóm. Chẳng hạn đối với nhóm thứ nhất:

$$S_1 = \sqrt{\frac{\sum_{j=1}^n (x_j - \bar{x})^2}{n-1}} = \sqrt{\frac{(15.4-14)^2 + \dots + (13.8-14)^2}{4-1}} = 1.11$$

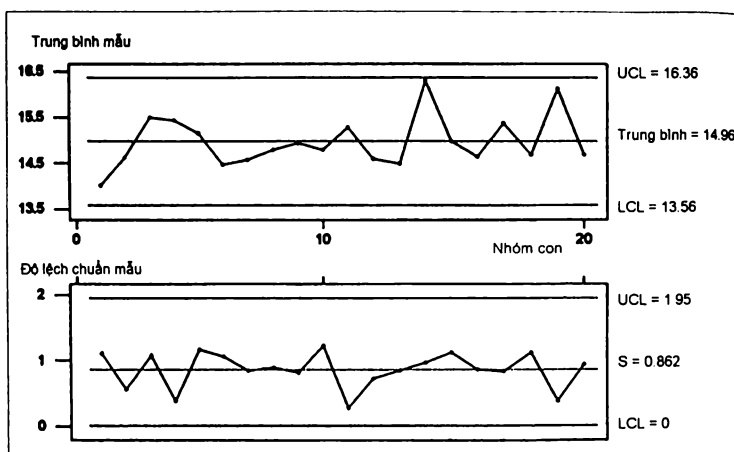
$$\bar{S} = \frac{\sum_{i=1}^m S_i}{m} = \frac{1.11 + 0.59 + \dots + 0.94}{20} = 0.862$$

Bảng 5.4. Dữ liệu độ dày bản mạch in với trung bình mẫu và độ lệch chuẩn mẫu

Nhóm con i	Độ dày				$\bar{x}_i$	S_i
1	15.4	14.1	12.7	13.8	14.000	1.11
2	15.3	14.2	14.1	14.8	14.600	0.56
3	15.5	15.0	17.0	14.5	15.500	1.08
4	15.7	15.1	15.8	15.1	15.425	0.38
5	14.7	16.8	14.1	15.0	15.150	1.16
6	14.3	15.8	13.2	14.5	14.450	1.07
7	14.3	15.8	14.2	13.9	14.550	0.85
8	14.4	15.3	15.7	13.7	14.775	0.90
9	14.2	16.0	14.4	15.1	14.925	0.81
10	13.2	14.9	16.2	14.8	14.775	1.23
11	15.4	15.6	15.0	15.1	15.275	0.28
12	13.9	14.0	15.3	15.1	14.575	0.73
13	15.2	13.4	15.1	14.2	14.475	0.85
14	17.7	15.6	15.7	16.2	16.300	0.97
15	13.5	14.8	15.5	16.1	14.975	1.12
16	14.3	14.0	15.9	14.3	14.625	0.86
17	14.3	16.2	15.8	15.1	15.350	0.83
18	14.6	15.9	13.2	15.0	14.675	1.12
19	15.7	15.9	16.5	16.4	16.125	0.39
20	15.9	14.9	13.8	14.1	14.675	0.94

Các hạn chế điều khiển và đường trung bình đối với biểu đồ điều khiển $\bar{x}$ được tính theo công thức, với $A_3 = 1.628$ như trong bảng khi $n = 4$. $UCL = \bar{\bar{x}} + A_3 \bar{S} = 14.96 + 1.628 * 0.862 = 16.36$; $CL = \bar{\bar{x}} = 14.96$; $LCL = \bar{\bar{x}} - A_3 \bar{S} = 14.96 - 1.628 * 0.862 = 13.56$.

Các hạn chế điều khiển đối với biểu đồ S được tính, với $B_3 = 0$; $B_4 = 2.226$, như cho trong bảng khi $n = 4$. $UCL = B_4 \bar{S} = 2.266 * 0.862 = 1.95$; $CL = \bar{S} = 0.862$; $LCL = B_3 \bar{S} = 0 * 0.862 = 0$.



Hình 5.21. Đồ thị điều khiển S và $\bar{x}$

Các biểu đồ điều khiển đối với trung bình mẫu và độ lệch chuẩn mẫu thể hiện trong hình. Quá trình được điều khiển, do không có khuôn mẫu hay xu thế rõ ràng, và các điểm đều trong vùng điều khiển của cả hai biểu đồ. Nếu có điểm bất thường, cần loại bỏ và tính lại các thông số điều khiển.

Các biểu đồ điều khiển riêng: cũng có trường hợp chỉ có một quan sát; $n = 1$. Mẫu với kích thước $n = 1$ khi sản xuất chậm và đắt, và nếu cho kích thước mẫu tăng lên sẽ không thực tế. Biểu đồ điều khiển riêng đối với dữ liệu biến thiên là thích hợp cho trường hợp này.

Tuy nhiên, quá trình với hai biến điều khiển sẽ có ý nghĩa hơn một biến, trong việc giám sát quá trình.

Biểu đồ điều khiển riêng dùng *hạng di chuyển*¹ của hai mẫu liên tiếp để đánh giá biến thiên của quá trình. Hạng di chuyển là $MR_i = |x_i - x_{i-1}|$. Để minh họa, đối với m nhóm con kích thước $n = 1$, $m - 1$ hạng di chuyển được xác định $MR_2 = |x_2 - x_1|$, ..., $MR_m = |x_m - x_{m-1}|$. Hạng di chuyển trung bình $MR = (MR_2 + \dots + MR_m) / (m - 1)$, do chỉ $m - 1$ giá trị hạng trung bình được tính, tức không có

¹ MR Moving range: hạng di chuyển.

hạng di chuyển đối với nhóm con thứ nhất. Các biểu đồ điều khiển được thiết kế đối với quan sát đơn, tức biểu đồ $\bar{x}$, và hạng di chuyển của nhóm con, tức biểu đồ MR. Biểu đồ MR được dùng để giám sát biến thiên của quá trình.

Các hạn chế điều khiển và đường trung bình của biểu đồ điều khiển $\bar{x}$ là: $UCL = \bar{\bar{x}} + 3\overline{MR} / d_2$; $CL = \bar{\bar{x}}$; $LCL = \bar{\bar{x}} - 3\overline{MR} / d_2$; trong đó d_2 là hằng số, phụ thuộc vào số các quan sát dùng để tính hạng di chuyển đối với mỗi nhóm con. Giá trị d_2 và các kích thước mẫu khác có trong bảng. Đặc biệt, do $\overline{MR}_1 = |x_1 - x_{1-1}|$, $n = 2$ đối với các biểu đồ điều khiển riêng. Biểu đồ điều khiển riêng được kiến thiết theo cách vẽ các quan sát thực, x_i , các hạn chế điều khiển, và đường trung bình, ứng với thứ tự các nhóm con.

Các hạn chế điều khiển và đường trung bình đối với biểu đồ điều khiển hạng di chuyển là: $UCL = D_4\overline{MR}$; $CL = \overline{MR}$; $LCL = D_3\overline{MR}$; trong đó D_3 và D_4 phụ thuộc vào kích thước mẫu, $n = 2$, cho trong bảng. Biểu đồ điều khiển hạng di chuyển được kiến thiết nhờ vẽ $m - 1$ hạng di chuyển, các hạn chế điều khiển, đường trung bình, ứng với thứ tự nhóm con.



Hình 5.22. Chai nước uống

Vẽ điều khiển riêng: Người ta dùng máy đổ đầy các chai nước uống 20 lít và cân. Trọng lượng của 25 chai liên tiếp được thu thập như dữ liệu trong bảng. Kỹ sư muốn xác định quá trình đổ đầy chai có trong vòng điều khiển không.

Bảng 5.5. Dữ liệu về chai nước uống.

Chai	Trọng số x_i	Khoảng di chuyển	Chai	Trọng số x_i	Khoảng di chuyển
1	20.01	—	14	20.01	0.01
2	19.99	0.02	15	20.06	0.05
3	19.95	0.04	16	19.91	0.15
4	20.62	0.67	17	20.06	0.15
5	20.04	0.58	18	20.03	0.03
6	20.00	0.04	19	20.10	0.07
7	19.83	0.17	20	19.89	0.21
8	19.94	0.11	21	19.95	0.06
9	19.94	0.00	22	20.00	0.05
10	20.08	0.14	23	20.01	0.01
11	20.01	0.07	24	20.05	0.04
12	20.04	0.03	25	20.05	0.00
13	20.00	0.04			

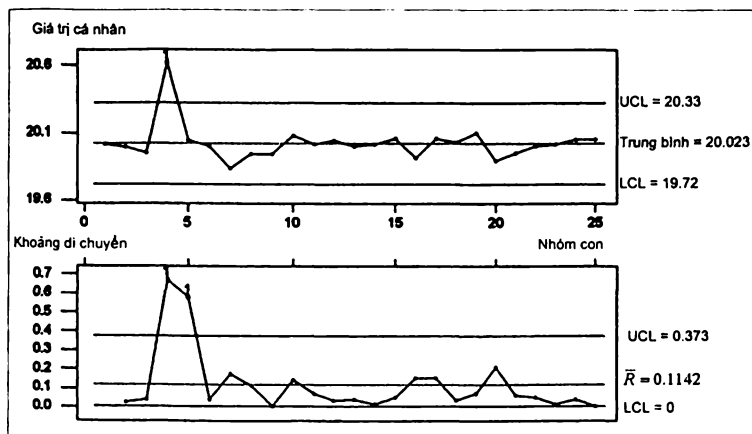
Giải pháp: Các hạng di chuyển được tính theo $MR_i = |x_i - x_{i-1}|$. Để minh họa việc tính, xét hạng di chuyển đầu tiên với nhóm con 2: $MR_2 = |x_2 - x_1| = |19.99 - 20.01| = 0.02$.

Các hạng di chuyển khác tính theo dữ liệu trong bảng. Để kiến thiết biểu đồ điều khiển đối với quan sát đơn, tức biểu đồ điều khiển $\bar{x}$, trọng lượng trung bình mẫu cho 25 chai được tính: $\bar{x} = (x_1 + \dots + x_m) / m = (20.01 + \dots + 20.05) / 25 = 20.023$.

Hạng di chuyển trung bình, $\overline{MR}$, được tính với các hạng di chuyển đối với 24 nhóm con. $\overline{MR} = (MR_2 + \dots + MR_m) / (m - 1) = ((0.02 + \dots + 0) = 0.1142$.

Các hạn chế điều khiển và đường trung bình đối với các biểu đồ riêng với các hạng di chuyển, kích thước 2: $UCL = \bar{x} + 3\overline{MR} / d_2 = 20.23 + 3 * 0.1142 / 1.128 = 20.33$; $CL = \bar{x} = 20.023$; $LCL = \bar{x} - 3\overline{MR} / d_2 = 20.023 - 3 * 0.1142 / 1.128 = 19.72$.

Các hạn chế điều khiển đối với biểu đồ hạng di chuyển với $D_3 = 0$ và $D_4 = 3.267$ từ bảng dữ liệu là $UCL = D_4\overline{MR} = 3.267 * 0.1142 = 0.373$; $CL = \overline{MR} = 0.1142$; $LCL = D_3\overline{MR} = 0 * 0.1142$.

Hình 5.23. Đồ thị điều khiển MR và $\bar{x}$

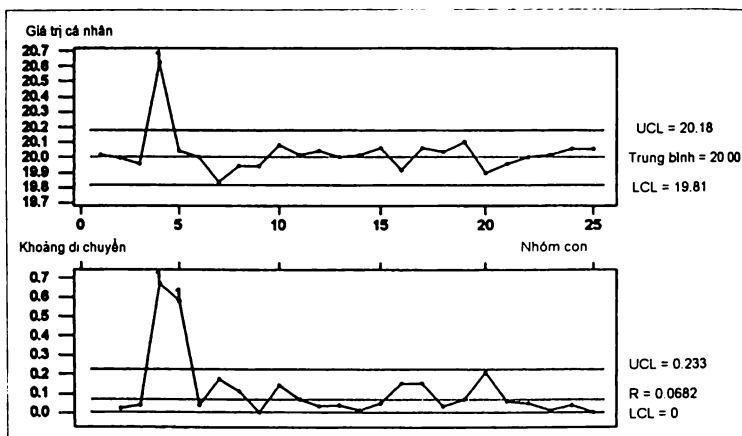
Các biểu đồ điều khiển đối với các quan sát đơn và đối với hạng di chuyển thể hiện trên hình. Có một vài điểm ra ngoài các hạn chế điều khiển trong cả biểu đồ riêng, lẫn biểu đồ về hạng di chuyển. Các điểm của nhóm con 4 trên biểu đồ riêng và nhóm con 4, 5 trên biểu đồ hạng di chuyển là bất thường. Vậy cần xét nhóm 4 và 5; nếu chúng thực sự là bất thường, cần loại bỏ.

Để minh họa cách xét lại biểu đồ điều khiển, giả sử nhóm con 4 và 5 là có lỗi, cần bị loại. Sau khi loại bỏ, cần tính lại trung bình mẫu, hạng di chuyển trung bình, các hạn chế điều khiển đối với cả hai biểu đồ. Bỏ quan sát 4, $x_4 = 20.62$, và quan sát 5, $x_5 = 20.04$, còn 23 nhóm con. Lấy m^* là số các nhóm con còn lại; $m^* = 23$.

Tính lại theo m^* , thay vì m , $\bar{x} = 20.00$; $MR = 0.0682$; $UCL^* = 3.267 * 0.0682 = 0.223$; $CL^* = 0.0682$; $LCL^* = 0 * 0.0682 = 0$.

Các biểu đồ điều khiển mới thể hiện trong hình. Nhận thấy điểm 4, 5 đã bị loại bỏ để tính lại các hạn chế điều khiển, được thể hiện trên biểu đồ. Để giữ trật tự thời gian hay trật tự nhóm con, cần giữ các quan sát này trên đồ thị, nhưng ghi nhận chúng không tham gia tính toán. Nói cách khác, các quan sát bị loại bỏ chỉ còn giữ chỗ, chứ không tác động đến các hạn chế điều khiển hay diễn

giải của biểu đồ. Các biểu đồ điều khiển mới không hiện ra bất kỳ quan sát mới, nằm ngoài các hạn chế điều khiển. Do vậy, quá trình xem như được điều khiển.



Hình 5.24. Đồ thị điều khiển ngược, ứng với thí dụ hình trên

Điều quan trọng là một mình biểu đồ điều khiển hạng di chuyển không diễn tả theo cách như biểu đồ R, khi xét khuôn mẫu hay xu thế. Một khuôn mẫu hay xu thế trên biểu đồ hạng di chuyển không nhất thiết là dấu hiệu về việc quá trình ngoài tầm điều khiển. Các hạng di chuyển có tương quan, gọi là $MR_i = |x_i - x_{i-1}|$. Có nhiều phụ thuộc giữa các quan sát kế tiếp tùy theo trật tự thời gian hay trật tự lập nhóm con. Trong trường hợp này, các khuôn mẫu hay xu thế có thể xảy ra tự nhiên.

5.4.2.2. Các biểu đồ điều khiển thuộc tính

Có nhiều quá trình hay các hoạt động thu thập dữ liệu không yêu cầu phép đo thực trên một phần cụ thể nào đó, như độ dài, rộng, trọng lượng... Có nhiều kịch bản mà đặc tính chất lượng đơn giản là phân lớp việc đo thành phạm trù đơn. Chẳng hạn các sản phẩm sản xuất ra được đo theo cách phân lớp hỏng/không hỏng, phù hợp/không phù hợp... Các tình huống khác có thể yêu

cầu tính những phần không phù hợp theo sản phẩm hay phạm trù. Chẳng hạn thanh thép được sản xuất được đánh giá tìm lỗi theo độ cứng, răng cưa... Trong các trường hợp này, các biểu đồ điều khiển biến là không cần thiết hay không còn phù hợp. Khi dữ liệu có thể được phân lớp, các biểu đồ điều khiển phù hợp được gọi là *biểu đồ điều khiển thuộc tính*.

Các biểu đồ điều khiển không phù hợp phần nhỏ: có hai biểu đồ không phù hợp phần nhỏ Shewhart (i) biểu đồ p ; (ii) biểu đồ np ¹. Thực tế hai biểu đồ cho kết quả như nhau, điều khiển vào và điều khiển ra, đối với bài toán cụ thể. Việc chọn biểu đồ để thực hiện là tùy ý và theo người thực hiện.

Trước hết cần thể hiện hình thức vài khái niệm về biểu đồ điều khiển. Đối với các biểu đồ điều khiển phần nhỏ, đặc tính chất lượng quan tâm được đặt chính xác vào một trong hai phạm trù. Các phạm trù này có thể là được/lỗi, hỏng/không hỏng, đúng/sai... Để đơn giản sử dụng thuật ngữ không phù hợp.

Định nghĩa: Không phù hợp có nghĩa là không đáp ứng, bị sai, mắc lỗi...

Các thông số sử dụng sau đây là:

- n : số các hạng mục được xem xét, kích thước lô hàng, hay kích thước mẫu;
- m : số các nhóm con được chọn;
- X : số cái không phù hợp thấy trong mẫu kích thước n , $X < n$;
- p : xác suất để hạng mục bất kì là không phù hợp;
- $\hat{p}$: *không phù hợp phần nhỏ mẫu*² thấy trong dữ liệu. Xác định $\hat{p} = X / n$, được tính đối với mỗi nhóm con;
- $\bar{p}$: *không thích hợp trung bình mẫu*. Xác định $\bar{p} = (\hat{p}_1 + \dots + \hat{p}_m) / m$, và là đánh giá của p .

¹ p , np : conform, nonconform: phù hợp, không phù hợp.

² Sample fraction nonconforming: không phù hợp phần nhỏ mẫu.

Đích của biểu đồ điều khiển không phù hợp phần nhỏ là giám sát tỉ lệ của hạng mục không phù hợp, đối với quá trình quan tâm, sử dụng dữ liệu thu thập trên m mẫu; mỗi mẫu có kích thước n .

Định nghĩa: Biểu đồ p là kĩ thuật biểu đồ dùng để giám sát tỉ lệ không phù hợp một cách trực tiếp.

Các hạn chế điều khiển và đường trung bình, khi chưa biết p , là: $CL = \bar{p}$;

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Các hạn chế điều khiển, đường trung bình, và điểm không phù hợp phần nhỏ mẫu $\hat{p}_i$ được vẽ theo số nhóm con, m . Nếu một hay nhiều điểm không phù hợp phần nhỏ nằm ngoài các hạn chế điều khiển, thì quá trình bị coi là ngoài điều khiển. Các khuôn mẫu và xu thế hiện ra cũng là chỉ thị về tình trạng mất điều khiển.

Bảng 5.6. Tài khoản ngân hàng bị lỗi, trong vòng 30 ngày

Ngày	Tài khoản có lỗi	$\hat{p}_i$	Ngày	Tài khoản có lỗi	$\hat{p}_i$
1	81	0.0405	16	75	0.0375
2	77	0.0385	17	73	0.0365
3	75	0.0375	18	75	0.0375
4	67	0.0335	19	80	0.0400
5	73	0.0365	20	62	0.0310
6	98	0.0490	21	76	0.0380
7	72	0.0360	22	78	0.0390
8	75	0.0375	23	72	0.0360
9	84	0.0420	24	89	0.0445
10	69	0.0345	25	82	0.0410
11	79	0.0395	26	82	0.0410
12	83	0.0415	27	91	0.0455
13	95	0.0475	28	62	0.0310
14	79	0.0395	29	79	0.0395
15	78	0.0390	30	76	0.0380

Vẽ giám sát tỉ lệ mắc lỗi: Xét ngân hàng địa phương thu thập dữ liệu về số hoạt động kế toán hàng ngày, để nhận biết sai sót. Các bản ghi về 2.000 hoạt động trong ngày, và dữ liệu được thu thập

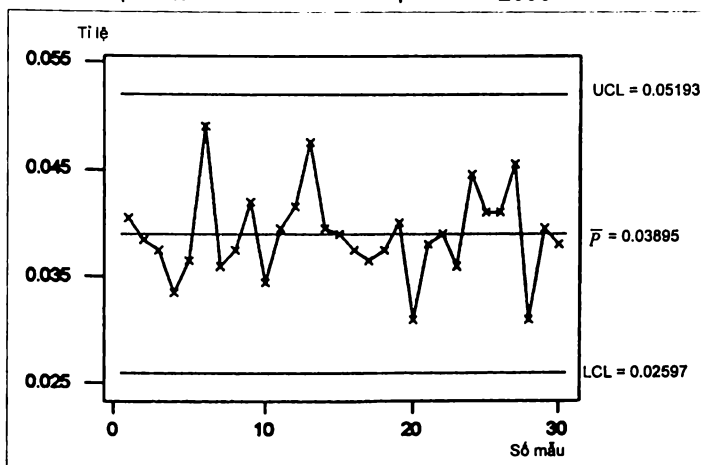
mỗi kì 30 ngày. Số các hoạt động kế toán bị lỗi được ghi lại trong bảng dữ liệu. Ngân hàng này muốn giám sát tỉ lệ lỗi mắc phải bằng cách thiết lập biểu đồ điều khiển.

Giải pháp: không phù hợp phần nhỏ w_i cần tính đối với mỗi ngày. Xét ngày thứ nhất: các hoạt động sai $X = 81$, trong tổng các hoạt động nghiên cứu, tức kích thước mẫu, $n = 2.000$. Tính toán thu được $\hat{p}_1 = 81 / 2.000 = 0.0405$. Các giá trị khác được tính tương tự, như trong bảng. Sự không phù hợp trung bình, $\bar{p}$, được xác định $\bar{p} = (\hat{p}_1 + \dots + \hat{p}_m) / m = (0.0405 + \dots + 0.0380) / 20 = 0.03895$.

Các hạn chế và đường trung bình đối với biểu đồ không thích hợp là $CL = 0.03895$;

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.03895 + 3\sqrt{\frac{0.03895(1-0.03895)}{2000}} = 0.05193;$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.03895 - 3\sqrt{\frac{0.03895(1-0.03895)}{2000}} = 0.02597$$



Hình 5.25. Đồ thị điều khiển không phù hợp với phần nhỏ

Biểu đồ p thể hiện trong hình. Không có điểm nào nằm ngoài các hạn chế điều khiển, và cũng không có xu thế hay khuôn mẫu nào. Do vậy, quá trình được điều khiển thống kê.

Biểu đồ điều khiển np: biểu đồ điều khiển np là biến thể của biểu đồ p, ngoài việc người ta sử dụng số các hạng mục không phù hợp, chứ không phải phần nhỏ. Biểu đồ np và biểu đồ p cho cùng một thông tin; nếu biểu đồ p cho thấy tình trạng ngoài điều khiển, thì biểu đồ np với cùng dữ liệu sẽ cho cùng dấu hiệu như vậy.

Một trong các lí do người ta sử dụng biểu đồ np thay cho biểu đồ p là thuận tiện cho kĩ thuật viên giải thích, trong lúc thu thập dữ liệu hay báo cáo kết quả.

Dưới khía cạnh thống kê, người ta cần biểu đồ này bởi lẽ sự không phù hợp phần nhỏ trung bình, tức $\bar{p}$, là giá trị duy nhất cần tính trước khi kiến thiết các hạn chế điều khiển. Không phù hợp phần nhỏ trung bình có thể thu được mà không cần tính $\hat{p}$. Đối với biểu đồ np, không phù hợp phần nhỏ trung bình được tính theo $\bar{p} = (X_1 + \dots + X_m) / (m \cdot n)$. Các hạn chế điều khiển được tính theo: $CL = n \bar{p}$;

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}; LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

Các hạn chế điều khiển, đường trung bình và số các hạng mục không phù hợp, X_i , được vẽ theo các nhóm con. Việc giải thích về biểu đồ này tương tự như với biểu đồ p.

Về biểu đồ np và không phù hợp toàn phần: xét bài toán trong thí dụ trước, thay vì kiến thiết biểu đồ p đối với không phù hợp phần nhỏ, người ta kiến thiết biểu đồ np đối với số các hạng mục không phù hợp.

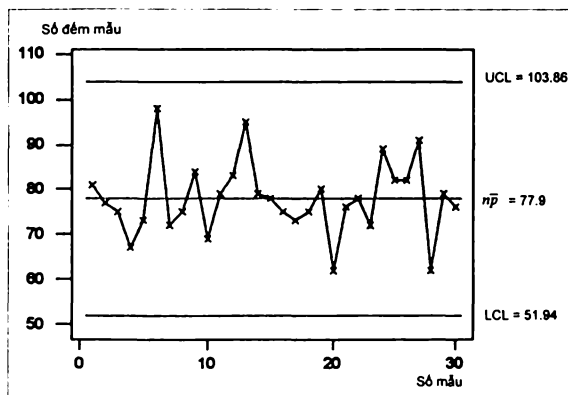
Đã cho $n = 2.000$, tính trong vòng 30 ngày, tức $m = 30$. Không phù hợp trung bình mẫu như trong thí dụ trước là $\bar{p} = 0.03895$. Tuy nhiên đối với thí dụ này, $\bar{p}$ được tính: $\bar{p} = (X_1 + \dots + X_m) / m / n = (81 + \dots + 76) / 30 / 2000 = 0.03895$, cùng giá trị như thí dụ trước.

Các hạn chế điều khiển được tính: $CL = n \bar{p} = 77.9$;

$$\begin{aligned} UCL &= n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} \\ &= 2000 \cdot 0.03895 + 3\sqrt{2000 \cdot 0.03895(1-0.03895)} = 103.86; \end{aligned}$$

$$LCL = \bar{np} - 3\sqrt{\bar{np}(1 - \bar{p})}$$

$$= 2000 * 0.03895 - 3\sqrt{2000 * 0.03895(1 - 0.03895)} = 51.94$$



Hình 5.26. Đồ thị điều khiển np đối với dữ liệu hoạt động của tài khoản

Biểu đồ thể hiện trong hình. Như với thí dụ trước, quá trình là được điều khiển thống kê.

Các biểu đồ điều khiển tính không phù hợp: các biểu đồ điều khiển về tính không phù hợp cũng tương tự như các biểu đồ đã trình bày, đối với số các hạng mục không phù hợp. Nhưng dưới góc nhìn của thuộc tính thống kê, chúng là khác nhau, và được tạo nên bởi các kĩ thuật khác nhau. Minh họa cho sự khác nhau này: (i) số các hạng mục không phù hợp X không vượt quá số các hạng mục nghiên cứu mẫu, n , tức $X < n$; (ii) để giám sát tính không phù hợp, không có hạn chế này. Tức tính không phù hợp được tính theo đơn vị; có hữu hạn không phù hợp trên đơn vị này, hay trên đơn vị nghiên cứu.

Chẳng hạn xét quá trình sản xuất chai nước. Muối chai được chọn mỗi giờ, để kiểm tra lỗi. Trong một mẫu 10 chai, người ta thấy có 17 lỗi. Trong kịch bản này, có 10 đơn vị nghiên cứu và phát hiện tính không phù hợp. Lưu ý rằng tính điều không phù hợp trong một mẫu. Có một số tình huống, chỉ chọn một đơn vị và ghi nhận những điều không phù hợp.

Biểu đồ điều khiển c: biểu đồ c là kỹ thuật biểu đồ điều khiển phù hợp đối với tính không phù hợp nếu kích thước mẫu, n , là không đổi giữa các nhóm con. Với biểu đồ c , người ta dùng các ký hiệu:

- n : số các hạng mục nghiên cứu, kích thước mẫu;
- m : số các nhóm con;
- X : số các điều không phù hợp trên hạng mục nghiên cứu, hay trên kích thước mẫu;
- $\bar{c}$: số trung bình các điều không phù hợp, trên toàn bộ bộ dữ liệu.

Theo định nghĩa, $\bar{c} = (X_1 + \dots + X_m) / m$.

Các hạn chế điều khiển được tính $UCL = \bar{c} + 3\sqrt{\bar{c}}$; $CL = \bar{c}$; $LCL = \bar{c} - 3\sqrt{\bar{c}}$.

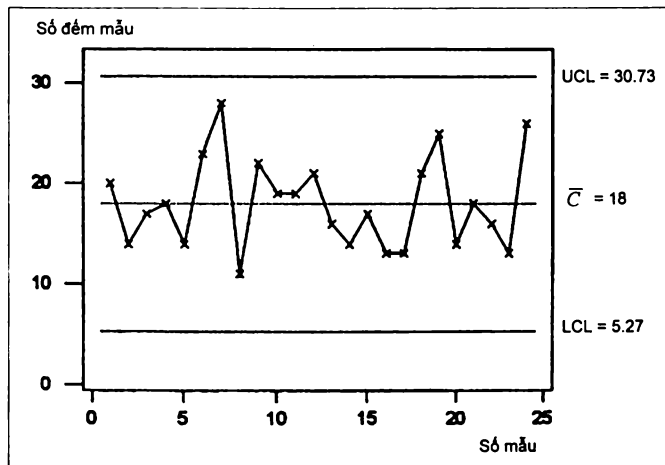
Về tính không phù hợp: xét quá trình sản xuất chai nước uống. Mười chai được chọn mỗi giờ, được kiểm tra về lỗi. Hai mươi tư mẫu như vậy được thu thập, ghi lại số các lỗi không phù hợp. Các kết quả thể hiện trong bảng. Kỹ sư muốn xác định về tính phù hợp của quá trình.

Bảng 5.7. Dữ liệu về quá trình sản xuất chai nhựa

Nhóm	Số lỗi	Nhóm	Số lỗi
1	20	13	16
2	14	14	14
3	17	15	17
4	18	16	13
5	14	17	13
6	23	18	21
7	28	19	25
8	11	20	14
9	22	21	18
10	19	22	16
11	19	23	13
12	21	24	26

Giải pháp: biểu đồ c là kỹ thuật giám sát thích hợp cho bài toán, do kích thước mẫu, 10 chai, là không đổi đối với các nhóm con. Số trung bình các lỗi không phù hợp là: $\bar{c} = (X_1 + \dots + X_m) / m = (20 + \dots + 26) / 24 = 18$.

Các hạn chế điều khiển và đường trung bình đối với số các lỗi không phù hợp là $UCL = \bar{c} + 3\sqrt{\bar{c}} = 18 + 3\sqrt{18} = 30.73$; $CL = \bar{c} = 18$; $LCL = \bar{c} - 3\sqrt{\bar{c}} = 18 - 3\sqrt{18} = 5.27$.



Hình 5.27. Đồ thị c đối với quá trình sản xuất chai nhựa

Biểu đồ điều khiển đối với quá trình thể hiện trong hình. Quá trình không thể hiện bất kì tình huống mất điều khiển nào. Quá trình này được chấp thuận theo dữ liệu thu thập.

Biểu đồ điều khiển u : biểu đồ u là kỹ thuật biểu đồ khác, liên quan đến tính không phù hợp. Biểu đồ u giám sát số trung bình của các lỗi không phù hợp; khác với biểu đồ c nhằm vào số các lỗi. Cũng như với biểu đồ n và np , các biểu đồ c và biểu đồ u cho cùng kết quả. Đối với biểu đồ này, cần bổ sung các kí hiệu:

- u : số trung bình của các lỗi không phù hợp trong một mẫu có n đơn vị. Theo định nghĩa $u_i = X_i / n$, tức số trung bình trên nhóm con i ;

- $\bar{u}$: số trung bình về không phù hợp theo đơn vị; $\bar{u} = (u_1 + \dots + u_m) / m$, là số trung bình trên tất cả các nhóm con.

Các hạn chế điều khiển và đường trung bình là; $CL = \bar{u}$;

$$UCL = \bar{u} + 3\sqrt{\frac{\bar{u}}{n}}; LCL = \bar{u} - 3\sqrt{\frac{\bar{u}}{n}}$$

Các hạn chế điều khiển và u_i được vẽ trên biểu đồ điều khiển theo nhóm con.

Về biểu đồ u về tính không phù hợp: lại xét bài toán nêu trong thí dụ trước. Số các chai được kiểm tra, số các lỗi được ghi lại, và số trung bình các lỗi theo chai, u_i , được liệt kê trong bảng sau.

Bảng 5.8. Dữ liệu về quá trình sản xuất chai nhựa đối với thí dụ trên

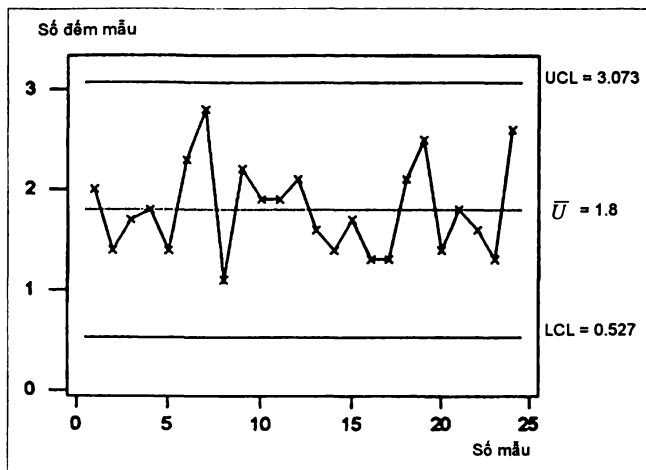
Nhóm	Số các lỗi	u_i	Nhóm	Số các lỗi	u_i
1	20	2.0	13	16	1.6
2	14	1.4	14	14	1.4
3	17	1.7	15	17	1.7
4	18	1.8	16	13	1.3
5	14	1.4	17	13	1.3
6	23	2.3	18	21	2.1
7	28	2.8	19	25	2.5
8	11	1.1	20	14	1.4
9	22	2.2	21	18	1.8
10	19	1.9	22	16	1.6
11	19	1.9	23	13	1.3
12	21	2.1	24	26	2.6

Minh họa cách tính u_i , xét số trung bình các lỗi không phù hợp đối với nhóm đầu tiên: $u_1 = X_1 / n = 20 / 10 = 2$. Số trung bình các lỗi theo chai, trên tất cả các nhóm con là $\bar{u} = (u_1 + \dots + u_m) / m = (2 + \dots + 2.6) / 24 = 1.8$.

Các hạn chế điều khiển là: $CL = \bar{u} = 1.8$;

$$UCL = \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} = 1.8 + 3\sqrt{\frac{1.8}{10}} = 3.073;$$

$$LCL = \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} = 1.8 - 3\sqrt{\frac{1.8}{10}} = 0.527$$



Hình 5.28. Đồ thị u đối với quá trình sản xuất chai nhựa

Các hạn chế điều khiển, đường trung bình và các giá trị u_i được vẽ theo mỗi nhóm con. Kết quả trên hình cho thấy không có dấu hiệu về mất điều khiển của quá trình.

Nhận xét về các biểu đồ điều khiển thuộc tính: các biểu đồ điều khiển thuộc tính là biểu đồ cơ bản và thông dụng nhất trong điều khiển quá trình thống kê. Mỗi một biểu đồ trong bốn biểu đồ có thể được phát triển với tình huống kích thước mẫu, n , thay đổi giữa các nhóm con.

Có thể chọn các kỹ thuật chuẩn để xử lý việc thay đổi của kích thước mẫu. (i) Người ta vẫn sử dụng các hạn chế điều khiển khi kích thước mẫu thay đổi. (ii) Hoặc người ta phát hiện kích thước mẫu trung bình đối với bài toán cụ thể, rồi áp dụng như trường hợp trong các thí dụ.

5.4.2.3. Biểu đồ điều khiển tổng tích lũy

Phần trên đã cho thấy các vai trò của biểu đồ Shewhart trong việc giám sát các quá trình. Các biểu đồ này là phù hợp nếu có giả thiết về tính chuẩn, đối với các đặc tính chất lượng quan tâm.

Đối với nhiều bài toán, giả thiết chuẩn không được đáp ứng; khi đó biểu đồ Shewhart đưa ra dấu hiệu không mạnh về các dịch chuyển nhỏ trong trung bình các quá trình, do chúng chỉ trả lời trên các *quan sát hiện tại*. Người ta thay các biểu đồ điều khiển Shewhart bằng *biểu đồ điều khiển tổng tích lũy*, CUSUM¹.

Biểu đồ CUSUM nhạy cảm nhiều với chuyển dịch nhỏ trong quá trình, do nó dựa vào cả dữ liệu giám sát, lẫn các quan sát quá khứ gần nhất. Nó được Page đề xuất năm 1954, liên quan đến thông tin từ các quan sát mẫu liên tiếp. Biểu đồ điều khiển về tổng tích lũy của các độ lệch theo giá trị đích nào đó. Biểu đồ điều khiển CUSUM vẽ các điểm định lượng $CS_i = (\bar{x}_1 - \mu_0) + \dots + (\bar{x}_n - \mu_0)$ đối với mỗi nhóm i ; trong đó:

- CS_i : tổng tích lũy cho đến và gồm mẫu thứ i ;
- $\bar{x}_j$: trung bình của mẫu j ;
- μ_0 : giá trị đích đối với trung bình quá trình;
- n : kích thước mẫu.

Khi quá trình đang trong trạng thái điều khiển được, tại giá trị đích μ_0 , tổng tích lũy CS_i xấp xỉ 0. Ngược lại, quá trình dịch ra khỏi trung bình đích, giá trị tuyệt đối của CS_i sẽ tăng nhiều. Tức có dịch chuyển âm và dương. Do CUSUM sử dụng thông tin từ dãy các quan sát, có thể phát hiện dịch chuyển nhỏ trong quá trình một cách nhanh chóng, so với biểu đồ Shewhart.

Biểu đồ điều khiển CUSUM có thể thực hiện với các nhóm con, hay với dữ liệu riêng. Biểu đồ điều khiển CUSUM hai phía đối với dữ liệu riêng sẽ được trình bày. Biểu đồ CUSUM một phía có thể được kiến thiết, ứng với hướng dịch chuyển cụ thể của dữ liệu.

Dạng *bảng*² của CUSUM yêu cầu hai thống kê (i) CS_i^+ và CS_i^- . CS_i^+ là tích lũy của độ lệch phía trên trung bình đích, được tham

¹ CUSUM: Cumulative sum control chart: biểu đồ tổng tích lũy.

² Tabular: bảng.

chiều như CUSUM về phía trên; CS_i^+ là tích lũy của độ lệch phía dưới trung bình đích, được tham chiếu như CUSUM về phía dưới. Chúng được tính: $CS_i^+ = \max\{0, x_i - (\mu_0 + K) + CS_{i-1}^+\}$; $CS_i^- = \max\{0, (\mu_0 + K) - x_i + CS_{i-1}^-\}$; với giá trị khởi đầu $CS_0^+ = CS_0^- = 0$, và x_i là quan sát thứ i . Hằng số K được gọi là *giá trị tham chiếu*, thường được tính theo $K = (|\mu_1 - \mu_0|) / 2$; trong đó μ_0 là trung bình đích, và μ_1 là trung bình mất điều khiển đang được quan tâm để phát hiện. Nếu trung bình quá trình nhận giá trị mới, người ta muốn phát hiện điều này sớm nhất.

Nếu chưa biết trung bình mất điều khiển, có các phương pháp xác định giá trị K . Khi đó lấy $K = k\sigma$, trong đó σ là độ lệch chuẩn của quá trình, và k là hằng số nào đó, để phát hiện được dịch chuyển cụ thể. Chẳng hạn người ta cho rằng dịch chuyển khỏi đích 1 độ lệch chuẩn là quan trọng, cần phát hiện; tức để phát hiện đích dịch chuyển hay không đến $\mu_0 + 1 * \sigma$ hay $\mu_0 - 1 * \sigma$, thì $k = 1$, và $K = 1 * \sigma$. Nếu độ lệch chuẩn của quá trình là không biết, người ta cần ước lượng từ dữ liệu đã có.

Biểu đồ điều khiển CUSUM hai phía vẽ các giá trị của CS_i^+ và CS_i^- đối với mỗi mẫu. Nếu điểm thống kê ra ngoài *khoảng quyết định*¹ H , quá trình này bị xem là mất điều khiển. Khi so sánh định lượng CS với H , cần lưu ý đến dấu âm, dương. Việc chọn H không phải ngẫu nhiên, mà chọn sau các xem xét cẩn thận. Thường $H = 5\sigma$ là khoảng quyết định chấp nhận được.

Về tổng tích lũy đối với độ lệch: ba mươi phép đo theo chiều được quan tâm, trong một phần của quá trình sản xuất. Tức kinh nghiệm đã qua với quá trình này, kĩ sư báo cáo về độ lệch chuẩn, trên chiều này, là $\sigma = 3.2$. Các phép đo ghi lại trên bảng. Đích của quá trình là $\mu_0 = 160$. Nếu dịch chuyển trung bình khỏi đích của quá trình này là nửa độ lệch chuẩn trung bình của quá trình, thì sản phẩm sẽ là không chấp nhận được. Kĩ sư muốn thiết kế biểu đồ điều khiển CUSUM hai phía và xác định quá trình trong tầm điều khiển không, với đích $\mu_0 = 160$.

¹ Decision Interval: khoảng quyết định.

Bảng 5.9. Dữ liệu về sản xuất

Quan sát	x_i	CS_i^+	CS_i^-	Quan sát	x_i	CS_i^+	CS_i^-
1	156.96	0	1.44	16	166.23	9.85	0
2	164.23	2.63	0	17	160.11	8.36	0
3	163.44	4.47	0	18	164.02	10.78	0
4	162.13	5	0	19	159.95	9.13	0
5	161.41	4.81	0	20	165.87	13.40	0
6	161.77	4.98	0	21	154.26	6.06	4.14
7	153.96	0	4.44	22	159.05	3.51	3.49
8	169.94	8.34	0	23	162.56	4.47	0
9	157.99	4.73	0.41	24	165.97	8.84	0
10	161.89	5.02	0	25	161.46	8.70	0
11	162.24	5.66	0	26	160.32	7.42	0
12	167.51	11.57	0	27	157.13	2.95	1.27
13	159.83	9.80	0	28	159.23	0.58	0.44
14	157.91	6.11	0.49	29	160.40	0	0
15	160.71	5.22	0	30	157.84	0	0.56

Giải pháp: Trước tiên, xác định vài con số cần thiết cho biểu đồ điều khiển. Theo bài toán $\mu_0 = 160$; $\sigma = 3.2$; $K = 0.5 * \sigma = 1.6$; do một nửa độ lệch chuẩn trung bình của quá trình là tham số cần phát hiện, $k = 1/2$; $K = k\sigma$; $H = 5\sigma = 16$.

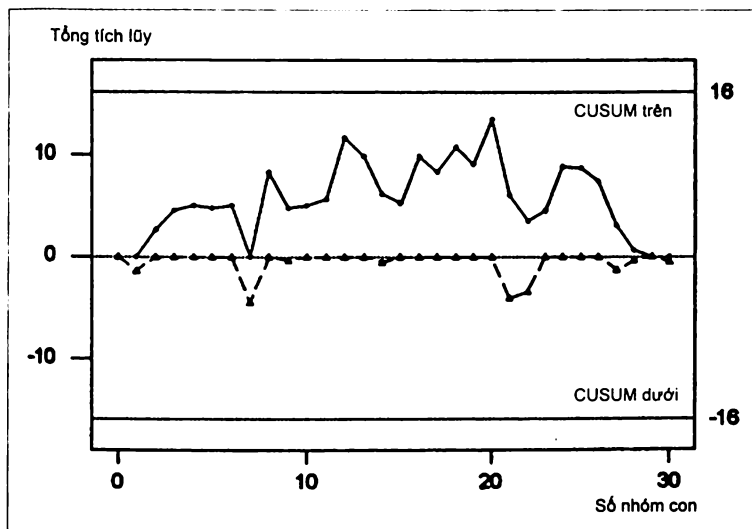
Các tổng tích lũy được tính đối với mỗi quan sát. $CS_i^+ = \max\{0, x_i - (\mu_0 + K) + CS_{i-1}^+\}$; $CS_i^- = \max\{0, (\mu_0 + K) - x_i + CS_{i-1}^-\}$. Về bản chất, nếu đích quá trình đang dịch chuyển, các tổng tích lũy sẽ khác 0. Sau khi tính các tổng tích lũy, đem chúng so sánh với khoảng quyết định $(-H, H) = (-16, 16)$. Nếu bất kì tổng tích lũy nào ra ngoài khoảng quyết định, quá trình này bị xem là mất điều khiển.

Minh họa với quan sát đầu tiên, $x_1 = 156.96$, với các giá trị khởi động $CS_0^+ = CS_0^- = 0$, $K = 1.6$, người ta thấy: $CS_1^+ = \max\{0, x_1 - (\mu_0 + K) + CS_0^+\} = \max\{0, 156.96 - (160 + 1.6) + 0\} = 0$; $CS_1^- = \max\{0, (\mu_0 + K) - x_1 + CS_0^-\} = \max\{0, (160 + 1.6) - 156.96 + 0\} = 1.44$. Do $CS_1^- = 1.44$, người ta sẽ so sánh 1.44 với khoảng quyết định.

Cả hai tổng tích lũy đều trong khoảng $(-H, H)$, tức $(-16, 16)$, nên quá trình này chưa cho thấy việc mất điều khiển.

Đối với quan sát thứ hai, $x_2 = 164.23$: $CS_2^+ = \max\{0, x_2 - (\mu_0 + K) + CS_1^+\} = \max\{0, 164.23 - (160 + 1.6) + 0\} = 2.63$; $CS_2^- = \max\{0, (\mu_0 + K) - x_2 + CS_1^-\} = \max\{0, (160 - 1.6) - 164.23 + 0\} = 0$. Cả hai tổng tích lũy đều trong khoảng quyết định.

Tương tự với các quan sát khác.



Hình 2.29. Đồ thị điều khiển CUSUM

Biểu đồ điều khiển SUCUM gồm các giá trị quyết định, các tổng tích lũy, với dấu giá trị, và đường trung bình tại 0, được thể hiện trong hình.

5.4.2.4. Biểu đồ điều khiển trung bình di chuyển có trọng số mũ

Biểu đồ điều khiển trung bình di chuyển có trọng số mũ EWMA¹ thay thế biểu đồ điều khiển Shewhart khi người ta chú trọng đến di chuyển nhỏ trong trung bình của quá trình. Biểu đồ

¹ EWMA: Exponentially Weighted Moving Average: trung bình di chuyển có trọng số mũ.

điều khiển EWMA được Roberts giới thiệu năm 1959. Thống kê EWMA được xác định $z_i = \lambda x_i + (1 - \lambda)z_{i-1}$; trong đó λ là trọng số, $0 < \lambda \leq 1$; x_i là quan sát hiện tại; z_{i-1} là thống kê EWMA trước, mà giá trị khởi đầu $z_0 = \mu_0$, tức trung bình đích của quá trình. Nếu chưa biết trung bình đích của quá trình, người ta sử dụng $\bar{x}$ làm giá trị khởi đầu.

Giá trị z_i có thông tin về quan sát quá khứ và quan sát hiện tại. Các hạn chế điều khiển áp lên giá trị z_i . Nếu một hay nhiều giá trị z_i ra khỏi hạn chế điều khiển, thì quá trình này bị xem là mất điều khiển. Các hạn chế điều khiển đối với EWMA với các giá trị i lớn là:

$$UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}}; LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}}$$

Trong đó L là độ rộng của các hạn chế điều khiển. Giá trị L và λ có thể tác động đáng kể đến hiệu quả của biểu đồ. Thông thường các giá trị λ nhỏ làm việc tốt, tức với $0.05 \leq \lambda \leq 0.25$. Giá trị L có xu hướng làm việc tốt khi $2.6 \leq L \leq 3$.

Về điều khiển trung bình di chuyển có trọng số mũ: xét dữ liệu về chiều như trong thí dụ 8 đối với biểu đồ CUSUM. Kỹ sư muốn kiến thiết biểu đồ điều khiển EWMA đối với dữ liệu này, sử dụng $\lambda = 0.1$ và $L = 2.7$. Các giá trị này không ngẫu nhiên, mà theo kinh nghiệm.

Giải pháp: như thí dụ trước, $\mu_0 = 160$, với $\sigma = 3.2$. Với $\lambda = 0.1$ và $L = 2.7$, thống kê EWMA được tính $z_i = \lambda x_i + (1 - \lambda)z_{i-1} = 0.1 * x_i + 0.9 * z_{i-1}$. Minh họa thống kê EWMA đối với mỗi quan sát với quan sát đầu tiên: $x_1 = 156.96$. Giá trị khởi đầu cho quá trình $z_0 = \mu_0 = 160$, z_1 được tính. Ở đây $z_1 = 0.1 * 156.96 + 0.9 * z_0 = 0.1 * 156.96 + 0.9 * 160 = 159.70$.

Các thống kê EWMA được tính tương tự; kết quả thể hiện trong bảng sau.

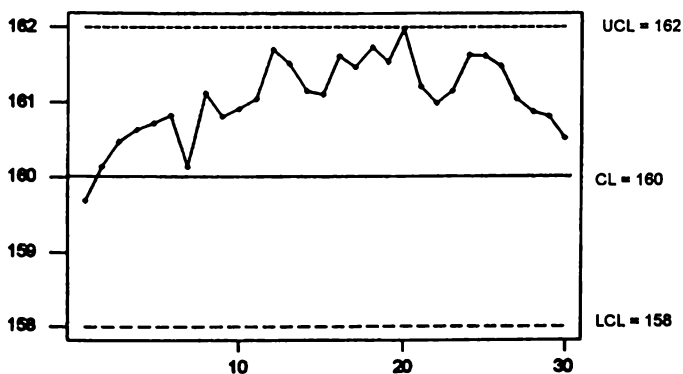
Bảng 5.10. Dữ liệu về máy

Số đo					
Quan sát	x_i	z_i	Quan sát	x_i	z_i
1	156.96	159.70	16	166.23	161.62
2	164.23	160.15	17	160.11	161.47
3	163.44	160.48	18	164.02	161.72
4	162.13	160.64	19	159.95	161.54
5	161.41	160.72	20	165.87	161.98
6	161.77	160.83	21	154.26	161.20
7	153.96	160.14	22	159.05	160.99
8	169.94	161.12	23	162.56	161.15
9	157.99	160.81	24	165.97	161.63
10	161.89	160.91	25	161.46	161.61
11	162.24	161.05	26	160.32	161.48
12	167.51	161.69	27	157.13	161.05
13	159.83	161.51	28	159.23	160.87
14	157.91	161.15	29	160.40	160.82
15	160.71	161.10	30	157.84	160.52

Các hạn chế điều khiển là: CL = 160;

$$UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}} = 160 + 2.7 * 3.2 * \sqrt{\frac{0.1}{2-0.1}} = 162;$$

$$LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}} = 160 - 2.7 * 3.2 * \sqrt{\frac{0.1}{2-0.1}} = 158$$



Các hạn chế điều khiển và các thống kê EWMA, z_1 , được vẽ trên biểu đồ điều khiển EWMA, như trong hình. Quá trình này trong tầm điều khiển, do tất cả các thống kê EWMA đều trong các hạn chế điều khiển.

5.4.2.5. Lựa chọn kĩ thuật biểu đồ điều khiển

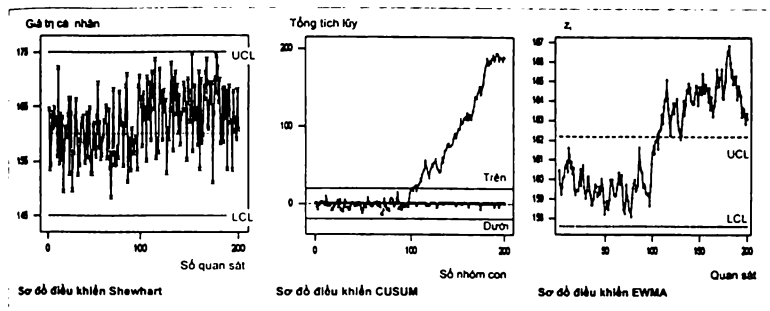
Phần trên đã đề cập biểu đồ Shewhart, biểu đồ CUSUM và biểu đồ EWMA. Quyết định phương pháp biểu đồ điều khiển nào tùy vào tình huống cụ thể trên quá trình được nghiên cứu.

- Các biểu đồ Shewhart thuận tiện cho việc kiến thiết và thực hiện, nhưng kém hiệu quả khi cần phát hiện các dịch chuyển nhỏ trong quá trình. Ngoài ra, biểu đồ điều khiển Shewhart quá nhạy với giả thiết về tính chuẩn. Nếu phân phối của quá trình không theo chuẩn, thì biểu đồ điều khiển Shewhart thường hiện dấu hiệu mất điều khiển, trái với thực tế;
- Các biểu đồ CUSUM và EWMA bền vững đối với xuất phát của dữ liệu về tính chuẩn, có xu hướng tốt hơn khi phát hiện chuyển dịch nhỏ trong trung bình quá trình.

Để minh họa cho sai khác về hiệu năng giữa biểu đồ điều khiển Shewhart và biểu đồ điều khiển CUSUM và EWMA, người ta đưa ra thí dụ sau.

Về so sánh các loại biểu đồ điều khiển: trong quá trình sản xuất cụ thể, dịch chuyển 0.5σ trong trung bình đích cần được phát hiện nhanh và nghiên cứu. Hai trăm mẫu về sản phẩm đã được chọn và có các biểu đồ. Để thấy được ưu việt của kĩ thuật giám sát CUSUM và EWMA so với kĩ thuật Shewhart, cho dịch chuyển trung bình của quá trình tại quan sát thứ 100. Một lược đồ giám sát tốt sẽ ra dấu về mất điều khiển sớm, sau khi thấy quan sát thứ 100.

Trong thí dụ này, người ta sử dụng cả ba kĩ thuật. Các biểu đồ điều khiển thể hiện trong hình.



Hình 5.31. Các đồ thị

Khi so sánh ba kĩ thuật, rõ ràng là kĩ thuật CUSUM và EWMA đã có dấu hiệu, trong hai quan sát, sau khi 0.5σ dịch chuyển xảy ra trong đích của quá trình.

5.4.2.6. Độ dài chạy trung bình

Độ dài chạy trung bình ARL^1 có thể dùng để đánh giá hiệu suất của biểu đồ điều khiển, hay xác định các giá trị thích hợp của các tham số đối với các biểu đồ đã trình bày. Chẳng hạn, để xác định các giá trị thích hợp nhất của λ và L đối với biểu đồ EWMA, người ta thiết lập giá trị chấp nhận được của ARL và xác định giá trị của λ và L đối với tình huống cụ thể.

Định nghĩa: Độ dài chạy trung bình ARL là số kì vọng của các mẫu, mẫu đã lấy trước khi xuất hiện dấu hiệu mất điều khiển trên biểu đồ điều khiển.

Nói chung, ARL là $ARL_0 = 1/p$; trong đó p là xác suất để điểm bất kì ra khỏi hạn chế điều khiển. Chẳng hạn xét biểu đồ điều khiển $\bar{x}$ với các hạn chế 3σ chuẩn. Các thí dụ trước cho thấy biểu đồ điều khiển Shewhart, $p = 0.0027$ là xác suất để một điểm trên biểu đồ điều khiển $\bar{x}$ ra ngoài hạn chế điều khiển, khi quá trình trong tầm điều khiển. Độ dài chạy trung bình trong tầm điều

¹ ARL : Average Run Length: độ dài chạy trung bình.

khiến đối với biểu đồ điều khiển $\bar{x}$ là $ARL_0 = 1/p = 1/0.0027 = 370$. Điều này có nghĩa dấu hiệu mất điều khiển là kì vọng xảy ra trong mỗi 370 mẫu, cho dù quá trình này điều khiển được về thực tế.

Hơn nữa, nếu quá trình thực sự mất điều khiển, người ta muốn giá trị ARL nhỏ. Giá trị ARL nhỏ cho thấy biểu đồ điều khiển sẽ ra dấu về mất điều khiển sớm, sau khi quá trình này dịch chuyển. ARL mất điều khiển là $ARL_1 = 1 / (1 - \beta)$; trong đó β là xác suất của việc không phát hiện dịch chuyển trên mẫu đầu tiên, sau khi đã xảy ra dịch chuyển. Nói khác đi, ARL_1 là số kì vọng của các mẫu đã quan sát, trước khi phát hiện thấy dịch chuyển.

Nói chung, ARL trong tầm điều khiển ứng với khái niệm về bảo động sai, khi mà ARL mất điều khiển có thể liên quan đến khái niệm về trễ khi phát hiện bất thường.

Các độ dài chạy trung bình được tính đối với mỗi kĩ thuật biểu đồ điều khiển. Việc tính này được thực hiện với tri thức về phân phối đã biết của dữ liệu và kích thước dịch chuyển cần phát hiện.

5.4.3. Các biểu đồ điều khiển nhiều biến

Có nhiều tình huống mà nhiều đặc trưng liên quan được giám sát đồng thời. sau đây trình bày hai kĩ thuật nhiều biến thông dụng (i) biểu đồ điều khiển T^2 ; (ii) biểu đồ điều khiển EWMA nhiều biến.

5.4.3.1. Mô tả dữ liệu

Khi hai hay nhiều đặc trưng chất lượng được nghiên cứu, hoàn toàn có thể có quan hệ nào đó, hay tương quan, giữa hai hay nhiều biến. Chẳng hạn đường kính trong và ngoài của turbin là hai đặc trưng cần được giám sát trong quá trình sản xuất.

Các biểu đồ điều khiển một biến được kiến thiết đối với mỗi đặc trưng chất lượng, như biểu đồ riêng, nhưng biểu đồ biến đơn không thể hiện được quan hệ giữa các biến. Thực tế, nếu có mỗi

quan hệ giữa các biến, các biểu đồ biến đơn sẽ đưa lại thông tin sai; tức là quá trình thực sự là mất điều khiển, nhưng biểu đồ một biến không phát hiện được. Hiện tượng này xảy ra khi có tương quan giữa các đặc trưng chất lượng, mà người ta lại nghiên cứu và giám sát chúng độc lập.

5.4.3.2. Biểu đồ điều khiển T^2

Biểu đồ điều khiển nhiều biến này do Hotelling đề xuất năm 1947. Điều khiển chất lượng nhiều biến có được nhiều quan tâm và có một số ứng dụng. Phân phối cơ bản của dữ liệu được giám sát theo biểu đồ điều khiển nhiều chiều. Phân phối chuẩn nhiều chiều đơn giản là phân nhiều biến, so với phân phối chuẩn đối với biến đơn.

Tự biểu đồ điều khiển T^2 có thể là phân đa biến so với biểu đồ điều khiển $\bar{x}$ Shewhart. Biểu đồ điều khiển này có thể thích nghi với dữ liệu nhóm con, hay với dữ liệu riêng.

Giả sử $x_1, x_2, \dots, x_p$ là p đặc trưng chất lượng, cần được giám sát đồng thời. Thống kê vẽ để giám sát là $T^2 = n (\bar{x} - \bar{\bar{x}})' S^{-1} (\bar{x} - \bar{\bar{x}})$; trong đó $\bar{x}$ là vector trung bình của mẫu; $\bar{\bar{x}}$ là vector của các trung bình trong điều khiển, được đánh giá từ dữ liệu quá trình; và S là ma trận hiệp phương sai trung bình đối với các đặc trưng chất lượng, khi quá trình trong tầm điều khiển. Trong trường hợp các quan sát riêng, $n = 1$, thống kê được vẽ là $T^2 = (x - \bar{x})' S^{-1} (x - \bar{x})$; trong đó $\bar{x}$ là vector trung bình mẫu; S là ma trận hiệp phương sai trung bình đối với các đặc trưng chất lượng, khi quá trình điều khiển được.

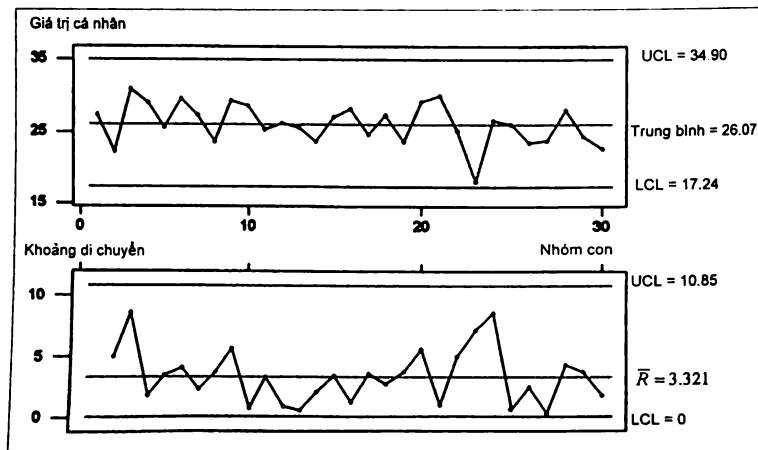
Xây dựng về các hạn chế và các thống kê vẽ ra là vấn đề chi tiết nên không trình bày trong phần này.

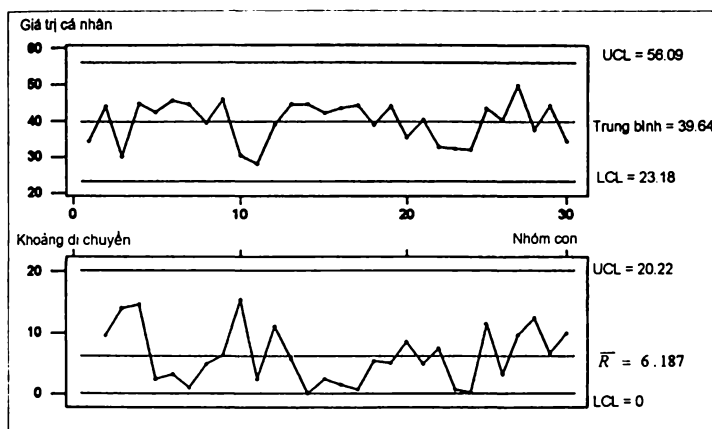
Về biểu đồ điều khiển nhiều biến: các phép đo trên các đường kích trong và ngoài turbin đồng cụ thể được thu thập và liệt kê trong bảng sau. Kỹ sư muốn xác định việc sản xuất có trong tầm điều khiển hay không.

Bảng 5.11. Dữ liệu về đường kính ống

Ống số	Đường kính trong x_1	Đường kính ngoài x_2	Ống số	Đường kính trong x_1	Đường kính ngoài x_2
1	27.2863	34.3854	16	28.1287	43.4413
2	22.2851	43.9282	17	24.5695	44.0917
3	30.8319	30.0322	18	27.2852	38.6781
4	29.0466	44.6336	19	23.5924	43.7138
5	25.6187	42.2535	20	29.1046	35.2320
6	29.6206	45.3383	21	30.0604	40.1509
7	27.3497	44.3806	22	25.1255	32.8442
8	23.6734	39.4534	23	18.0000	32.1295
9	29.3292	45.6672	24	26.5506	31.9827
10	28.6392	30.2690	25	25.9848	43.3962
11	25.3458	27.9130	26	23.4975	40.2245
12	26.1724	38.8171	27	23.7466	49.7419
13	25.6517	44.4404	28	28.0258	37.4064
14	23.6246	44.3296	29	24.3527	44.0640
15	26.9687	41.9942	30	22.5746	34.1724

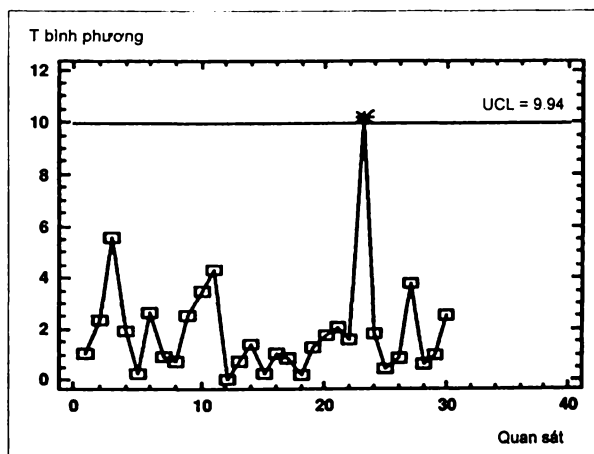
Giải pháp: Để xây dựng các biểu đồ điều khiển một biến, đối với mỗi đặc tính chất lượng. Đặc biệt, biểu đồ điều khiển Shewhart đã trình bày có thể sử dụng cho cả hai đặc tính chất lượng. Kết quả thấy trong các hình, dựa trên các biểu đồ điều khiển riêng, quá trình là điều khiển thống kê được.

**Hình 5.32. Đồ thị điều khiển MR và giá trị riêng Shewhart đối với đường kính trong ống**



Hình 5.33. Đồ thị điều khiển MR và giá trị riêng Shewhart đối với đường kính ngoài ống

Nếu chỉ xây dựng biểu đồ điều khiển một biến, quá trình được xem là điều khiển được. Nhưng khi xây dựng biểu đồ điều khiển T^2 , như trong hình trên, người ta thấy quá trình mất điều khiển do có quan sát ra ngoài hạn chế điều khiển trên. Sờ dĩ thu được kết quả về mặt điều khiển là do kĩ thuật T^2 của Hotelling chú trọng mối quan hệ giữa hai biến, như có một biến.



Hình 5.34. Đồ thị điều khiển T^2 Hotelling đối với đường kính trong, ngoài ống

Trong trường hợp này, biểu đồ điều khiển nhiều biến có nhiều thông tin hơn các biểu đồ điều khiển một biến đồng thời. Khi giám sát nhiều đặc trưng chất lượng, kĩ thuật biểu đồ điều khiển nhiều biến là phương pháp phù hợp nhất, để định ra quá trình mất điều khiển.

5.4.3.3. Biểu đồ điều khiển EWMA nhiều biến

Một trong các hạn chế của biểu đồ điều khiển T^2 của Hotelling, cũng như hạn chế của các biểu đồ Shewhart, là chi sử dụng quan sát hay thông tin về nhóm con hiện tại. Các biểu đồ này thường không nhạy với dịch chuyển nhỏ trong trung bình đích của quá trình. Người ta thay phương pháp T^2 bằng phương pháp tương tự như biểu đồ điều khiển EWMA một biến.

Biểu đồ điều khiển EWMA đối với đặc tính chất lượng đơn được mở rộng cho tình huống nhiều biến. Năm 1992, người ta phát triển biểu đồ điều khiển MEWMA¹, và người ta cũng trao đổi về các thuộc tính của ARL cho biểu đồ này. Thống kê MEWMA được xác định $Z_i = \lambda x_i + (1 - \lambda)Z_{i-1}$; trong đó Z và x là các vectơ nhiều chiều; tức x_i đại diện cho quan sát thứ i trong bộ dữ liệu. Chẳng hạn xét bài toán về đường kính turbin cho trong bảng, $x_1 = (27.2863, 34.3854)$, thể hiện ống đồng thứ nhất.

Đối với MEWMA, đại lượng vẽ trên biểu đồ được xác định $T_i^2 = Z_i' \Sigma_i^{-1} Z_i$; trong đó $\Sigma_i = (\lambda / (2 - \lambda)) (1 - (1 - \lambda)^{2i}) \Sigma$, là ma trận hiệp phương sai, là dạng tương tự nhiều biến của EWMA một biến.

Người ta đã cho bảng và các biểu đồ đối với các hạn chế điều khiển và các giá trị của nhân tố trọng số λ .

5.4.4. Nhận xét về xử lí dữ liệu bất thường

Quan trọng của giám sát và điều khiển quá trình luôn còn đó. Phần trên mới chỉ là một vài kĩ thuật biểu đồ điều khiển, dùng

¹ MEWMA: Multivariate EWMA: trung bình di chuyển có trọng số mũ nhiều chiều.

trong điều khiển quá trình và giám sát. Các kỹ thuật này dễ thích nghi để sử dụng khi phát hiện bất thường.

Các biểu đồ điều khiển Shewhart là hữu dụng trong tình huống mà dịch chuyển về trung bình của quá trình là nhỏ, ít quan trọng, không cần phát hiện nhanh như đối với các di chuyển lớn. Đối với dịch chuyển nhỏ về trung bình quá trình, các biểu đồ CUSUM và WEMA có hiệu quả hơn, với nghĩa độ dài chạy trung bình trong và ngoài điều khiển.

Biểu đồ điều khiển T^2 của Hotelling và MEWMA là tốt hơn với thực hiện đồng thời các biểu đồ điều khiển một biến, đối với quá trình cần đến hai hay nhiều biến.

5.5. Phân tích dữ liệu và không chắc chắn

5.5.1. Giới thiệu dữ liệu không chắc chắn

Dữ liệu không chắc chắn sinh ra khi khớp thế giới thực vào cơ sở dữ liệu, và thế giới thực cũng không rõ ràng. Công cụ cơ bản để xử lý dữ liệu không chắc chắn là xác suất. Vậy nên cần bắt đầu với việc kiến thiết các mô hình thống kê.

Nền tảng đối với nhiều hoạt động khai phá dữ liệu là kỹ pháp về mẫu¹. Đôi khi cơ sở dữ liệu chứa chỉ một mẫu từ không gian các bản ghi có thể có. Một số khía cạnh quan tâm là (i) khuôn mẫu; (ii) đánh giá mô hình mô tả dữ liệu, đặc biệt mô hình Bayes; (iii) đánh giá chất lượng của giả thuyết trên cơ sở các dữ liệu quan sát.

5.5.2. Xử lý không chắc chắn

Nghĩa rộng của không chắc chắn được miêu tả như sự thay đổi nhiều của thế giới dùng để mô tả và các khái niệm liên quan. Xác suất, cơ hội, tính ngẫu nhiên, may mắn, đột xuất và số phận là các thí dụ. Việc không chắc chắn có ở mọi nơi, yêu cầu người ta xem việc mô hình hóa chắc chắn là thành phần của phân tích dữ liệu.

¹ Sample: mẫu.

Không chắc chắn do vài nguyên nhân: dữ liệu chỉ lấy mẫu từ đám dân cư nghiên cứu, vậy người ta đã không chắc chắn về hướng mờ rộng sang mẫu khác, về toàn bộ dân cư. Có thể người ta đã suy đoán về tương lai trên dữ liệu hiện tại, đã phỏng đoán khi không thu thập được một số dữ liệu.

Nhiều cơ sở khái niệm được dùng để quản lí không chắc chắn, sự không biết. Hầu hết chúng đều sử dụng xác suất. Logic mờ cũng là một cách; nó liên hệ chặt với lí thuyết xác suất và tập thô; tuy nhiên thiếu nền lí thuyết để ứng dụng được rộng rãi như xác suất.

5.5.2.1. Lí thuyết xác suất, tính toán xác suất

Cần phân biệt lí thuyết xác suất với *tính toán xác suất*¹. Lí thuyết xác suất liên quan đến diễn giải của thống kê; tính toán xác suất liên quan đến xử lí các thể hiện toán học của xác suất.

Việc phân biệt này quan trọng ở chỗ cho phép tách các lĩnh vực, giữa các quy định tổng quát trong tính toán, với việc phân biệt các ý kiến trong lí thuyết. Theo Kolmogorov, 1930, tính toán là nhánh của toán học, dựa trên các tiên đề được công nhận và được xác định rõ, nhằm khai phá hệ quả của các tiên đề. Người ta dùng các hệ tiên đề khác nhau, quá đặc biệt và tổng quát, không dùng trong khai phá dữ liệu. Lí thuyết cho thấy viễn cảnh khớp thể giới thực với thể hiện toán học, là cái khai phá dữ liệu cần.

5.5.2.2. Khung nhìn xác suất

Một nghiên cứu về lịch sử triết học của lí thuyết xác suất cho thấy có nhiều viễn cảnh mang ý nghĩa của xác suất như các học giả đã nghĩ đến. Tuy nhiên người ta có thể nhóm lại vào một số dạng. Liên quan đến khai phá dữ liệu, người ta xét hai dạng quan trọng nhất (i) xác suất khách thể, với *khung nhìn thường xuyên*; (ii) xác suất chủ thể, với *khung nhìn cạnh tranh*².

¹ Theory: lí thuyết; Calculus: tính toán.

² Frequentist view: khung nhìn thường xuyên; Competing view: khung nhìn cạnh tranh.

1. *Khung nhìn thường xuyên* của xác suất cho biết xác suất là khái niệm khách thể. Cụ thể, xác suất của sự kiện được xác định như giới hạn của tỉ lệ xuất hiện lặp lại trong hoàn cảnh như nhau. Trong tất cả quá trình mô hình hóa khoa học, người ta cần xác định khía cạnh quan trọng cho mô hình để đủ chính xác;
2. Trong chục năm lại đây, *khung nhìn cạnh tranh* cho rằng xác suất không là đối tượng mà là chủ thể, khi hình thức hóa kí pháp xác suất. Nguyên tắc và phương pháp luận đối với phân tích dữ liệu theo quan điểm của xác suất chủ thể là thống kê Bayes. Thống kê Bayes đặc trưng hóa rõ về tất cả các dạng không chắc chắn trong bài toán phân tích dữ liệu. Sự không chắc chắn trong tập các cấu trúc mô hình là gần nhất với thực tế. Xác suất chủ thể là phần linh hoạt đối với việc mô hình hóa sự kiện không chắc chắn theo các dạng khác nhau.

Theo quan điểm của xác suất chủ thể, xác suất là mức độ tin cậy về sự xảy ra của sự kiện. Do vậy xác suất không là tính chất khách thể của thế giới bên ngoài, mà là trạng thái bên trong của cá thể, và dùng để phân biệt cá thể này với cá thể khác. May mắn là người ta chấp nhận một số nguyên lí cho xác suất chủ thể như với xác suất khách thể. Việc tính toán nêu trên cũng như đối với hai quan điểm, cho dù nó có sự diễn giải hoàn toàn khác.

Người ta không sử dụng hai tiếp cận đồng thời. Xác suất chủ thể có thể áp dụng được trong lĩnh vực mà khung nhìn thường xuyên với xác suất khách thể không áp dụng được. Ngoài ra, suy luận thống kê dựa trên xác suất chủ thể cần có thành phần chủ thể, tức độ tin cậy trước, ban đầu về việc xảy ra sự kiện. Nhân tố này khác nhau đối với mỗi người.

Đôi khi tiếp cận thường xuyên và tiếp cận chủ thể đều cho cùng kết quả; đó là khi dùng giả thuyết đơn giản và bộ dữ liệu lớn. Lẽ ra cần phân biệt tiếp cận của mình, có người chấp nhận kết quả theo tiếp cận mình nghĩ. Phương pháp luận về phân tích dữ liệu rút

ra từ tiếp cận thường xuyên có khuynh hướng đơn giản về tính toán, nên được dùng cho kĩ thuật khai phá dữ liệu mà bộ dữ liệu không phù hợp với tính toán phức tạp. Khi áp dụng các phương pháp Bayes, có thể thu được thêm thông tin. Trong mấy năm qua, thống kê quan tâm nhiều đến khai phá dữ liệu với tiếp cận Bayes. Phần tiếp theo sẽ trình bày theo cách hòa giải hai tiếp cận: phương pháp luận thường xuyên về làm khít các mô hình và các mẫu dữ liệu được thực hiện với trường hợp đặc biệt của phương pháp luận Bayes tổng quát. Đối với thực tế, điều này có ý nghĩa ở chỗ: cả mô hình tổng quát và phân tích toán được sử dụng.

5.5.3. Các biến ngẫu nhiên và các mối quan hệ

5.5.3.1. Biến ngẫu nhiên

Định nghĩa: Biến ngẫu nhiên là thể hiện khớp thuộc tính của các đối tượng với một biến có thể nhận một trong các giá trị có thể được, thông qua quá trình người quan sát thấy vài yếu tố chưa đoán trước được của đối tượng.

Các giá trị của biến ngẫu nhiên X tạo nên miền của X . Người ta dùng X để chỉ biến, x chỉ giá trị của biến ngẫu nhiên.

Thuộc tính cơ bản của các biến ngẫu nhiên, một chiều đơn, gồm cả các hàm xác suất $p(X)$, trong đó miền X hữu hạn, và hàm mật độ xác suất $f(x)$, trong đó miền X là thực tế hay khoảng xác định thực. Kỳ vọng của X : $E(x) = \int xf(x)dx$, đối với X có giá trị thực; do E là phép toán tuyến tính, $E[X + Y] = E[X] + E[Y]$.

5.5.3.2. Các biến ngẫu nhiên nhiều chiều

Do khai phá dữ liệu thường gặp biến nhiều chiều, người ta đưa ra khái niệm biến ngẫu nhiên nhiều chiều.

Định nghĩa: Biến ngẫu nhiên nhiều chiều X là tập các biến ngẫu nhiên $X_1, \dots, X_p$.

Dùng vector p chiều $x = \{x_1, \dots, x_p\}$ để kí hiệu tập các giá trị của X . Hàm mật độ $f(X)$ của biến ngẫu nhiên nhiều chiều X được gọi là

*hàm mật độ đồng thời*¹ của X . Kí hiệu hàm này bằng $f(X) = f(X_1 = x_1, \dots, X_p = x_p)$, hay đơn giản là $f(x_1, \dots, x_p)$. Các phân phối xác suất liên kết đối với các biến lấy các giá trị trong tập hữu hạn. Lưu ý hàm $f(X)$ là hàm vô hướng của p biến.

Hàm mật độ của biến đơn bất kì trong tập X , hay tổng quát của tập con bất kì của tập toàn thể các biến, được gọi là *hàm lẻ*² của mật độ đồng thời. Về kĩ thuật, nó được rút ra từ mật độ đồng thời bằng cách tóm tắt hay tích hợp qua các biến không gồm tập con. Chẳng hạn đối với biến ngẫu nhiên 3 chiều $X = (X_1, X_2, X_3)$, mật độ lẻ của $f(X_1)$ được lấy là $f(x_1) = \iint f(x_1, x_2, x_3) dx_2 dx_3$.

Mật độ của biến đơn khi cho, hay có điều kiện, các giá trị cụ thể của các biến khác là *mật độ điều kiện*³. Do vậy, có thể nói về mật độ điều kiện của biến X_1 khi X_2 lấy giá trị 6, kí hiệu $f(x_1 | x_2 = 6)$. Tổng quát, mật độ điều kiện của X_1 khi có giá trị của x_2 , kí hiệu $f(x_1 | x_2)$, được xác định $f(x_1 | x_2) = f(x_1, x_2) / f(x_2)$.

Đối với các biến ngẫu nhiên giá trị rời rạc, người ta dùng $p(x_1 | x_2) \dots$. Cũng có thể trộn cả hai, tức hàm mật độ xác suất điều kiện $f(x_1 | a_1)$ với biến liên tục có điều kiện theo biến phạm trù, với hàm xác suất có điều kiện $p(a_1 | x_1)$ đối với trường hợp ngược lại.

Độc lập, phụ thuộc: các biến trong tập X nhiều chiều có thể liên quan với nhau. Mà vấn đề tổng quát trong khai phá dữ liệu là phát hiện mối quan hệ giữa các biến.

Định nghĩa: Các biến được coi là độc lập nếu không có mối quan hệ trong các giá trị thể hiện của các biến; ngược lại chúng là phụ thuộc.

Một cách hình thức, các biến X và Y là độc lập chỉ nếu $p(x, y) = p(x) \cdot p(y)$ đối với tất cả các giá trị của X và Y . Dạng tương đương là $p(x | y) = p(x)$ hay $p(y | x) = p(y)$ đối với tất cả giá trị của X và Y . Đó là hàm xác suất; sau này dùng $f(x, y) = f(x) \cdot f(y)$. Định nghĩa về

¹ Joint density function: hàm mật độ đồng thời.

² Margin function: hàm lẻ.

³ Conditional density: mật độ điều kiện.

sự độc lập cho thấy khi X và Y là độc lập, phân phối của X không đổi dù Y có xuất hiện hay không. Do vậy Y không mang thông tin về X , theo nghĩa giá trị Y nhận không ảnh hưởng đến xác suất để X nhận giá trị bất kì.

Người ta có thể tổng quát hóa ý này cho nhiều hơn hai biến. Nói rằng X là độc lập có điều kiện của Y khi cho Z nếu đối với tất cả giá trị của X, Y và Z , người ta có $p(x, y | z) = p(x | z) \cdot p(y | z)$.

Lưu ý rằng độc lập có điều kiện không ngầm định độc lập lẽ (có điều kiện), tức quan hệ trên không ngầm định $p(x, y) = p(x) \cdot p(y)$.

Giả thiết về độc lập có điều kiện được dùng rộng rãi trong ngữ cảnh các dữ liệu dãy, trong đó giá trị tiếp theo trong dãy thường độc lập với tất cả các giá trị quá khứ trong dãy đã biết. Theo ngữ cảnh này, độc lập có điều kiện đã được xét với tên là tính chất Markov bậc một¹.

Kí pháp về độc lập và độc lập có điều kiện, như tổng quát hóa của độc lập, là trọng tâm của nhiều khái niệm quan trọng trong phân tích dữ liệu. Giả thiết độc lập và độc lập có điều kiện cho phép người ta xác định nhân tố của tỉ trọng liên kết của nhiều biến vào kết quả để dùng hơn với tỉ trọng đơn giản hơn, tức

$$f(x_1, \dots, x_n) = f(x_1) \cdot \prod_{j=2}^n f(x_j | x_{j-1}); \text{ trong đó mỗi biến } x_i \text{ là độc lập có}$$

điều kiện với $x_1, \dots, x_{i-2}$ khi cho giá trị x_i . Đây là thí dụ về mô hình Markov bậc một. Ngoài ưu điểm về tính toán do đơn giản hóa, người ta còn cấu trúc lại mô hình để sử dụng ít biến hơn. Thế nhưng giả thiết về độc lập hay bị vi phạm trong thực tế, chẳng hạn giả thiết về dãy các kí tự trong văn bản là dãy Markov bậc một là không thực tế. Luôn tâm niệm rằng các mô hình chỉ là xấp xỉ của thế giới thực, ưu điểm do giả thiết độc lập phù hợp sẽ tốt hơn ưu điểm do chọn được mô hình phức tạp, mà kém bền vững.

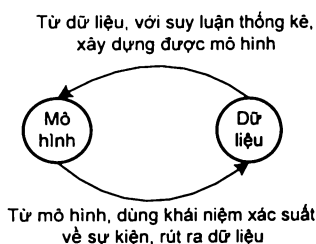
Một trường hợp đặc biệt của phụ thuộc là tương quan, hay phụ thuộc tuyến tính, đã nêu ra trước đây. Lưu ý phụ thuộc thống

¹ The first - order Markov: Markov bậc một.

kê không giống như tương quan: hai biến có thể phụ thuộc nhưng không có tương quan tuyến tính. Các biến được xem là có tương quan dương nếu giá trị lớn của biến này liên quan với giá trị cao biến kia, tương quan âm nếu giá trị cao biến này tương quan với giá trị thấp biến kia. Không được lẫn tương quan với nhân quả. Hai biến có thể là tương quan dương, mà không có chút quan hệ nhân quả nào.

5.5.4. Các mẫu và suy luận thống kê

Nhiều bài toán khai phá dữ liệu yêu cầu toàn bộ dân cư, nhiều bài toán khác chỉ cần dữ liệu lấy làm *mẫu*¹ của dân cư. Đối với các bài toán lấy dữ liệu làm mẫu, các mẫu hiện ngay lúc ban đầu. Cũng có khi cho dù toàn dân cư sẵn sàng, các thao tác chỉ thực hiện trên các mẫu. Lúc này dữ liệu vẫn hợp lệ, dù bài toán có đích là mô hình, cần chỉ ra cấu trúc dữ liệu cho tất cả. Các cấu trúc này cũng thể hiện trên dữ liệu mẫu, miễn là dữ liệu không ít quá. Tất nhiên làm việc với mẫu nhỏ của bộ dữ liệu lớn sẽ không phù hợp, nếu đích bài toán là phát hiện khuôn mẫu², tức đích là phát hiện sự lệch lạc nhỏ so với đồng dữ liệu, và nếu ít dữ liệu sẽ khó thấy sự lệch này. Ngoài ra, nếu đích là phát hiện các bản ghi có hành vi bất thường, các nhà phân tích cần dựa trên mẫu toàn thể.



Hình 5.35. Liên quan đến mô hình và xác suất

Chính lúc một mẫu được dùng là cần đến suy luận thống kê. Suy luận thống kê cho phép khẳng định về các cấu trúc dân cư,

¹ Sample: mẫu.

² Pattern detection: phát hiện khuôn mẫu

đánh giá kích thước của các cấu trúc này, và thiết lập mức tin cậy vào chúng, tất cả trên cơ sở của một mẫu. Thuật ngữ “đánh giá” vừa dùng là cho giá trị dân cư. Nếu dựa trên phân tích trên toàn dân cư, người ta sẽ dùng “tính”: nếu biết tất cả các giá trị tạo nên, người ta có thể tính giá trị dân cư, không phải “đánh giá”.

Để suy luận về cấu trúc dân cư, cần có cấu trúc *mô hình* và *khuôn mẫu*¹; không thể thêm gì vào cấu trúc mà chưa rõ cấu trúc đó.

Suy luận thống kê dựa trên giả thiết rằng mẫu được rút ra từ dân cư theo cách ngẫu nhiên; mỗi phần tử của dân cư có xác suất cụ thể về khả năng xuất hiện trong mẫu. Mô hình sẽ xác định hàm phân phối cho dân cư này, tức xác suất để một giá trị cụ thể của biến bất kì sẽ xuất hiện trong mẫu. Cho mô hình, nói chung người ta có thể tính được xác suất để quan sát rơi vào một khoảng. Đối với các mẫu của phân phối phạm trù, người ta có thể đánh giá xác suất để giá trị bằng với các giá trị quan sát sẽ xảy ra. Tổng quát, khi người ta có mô hình M của dữ liệu, người ta có thể thiết lập xác suất để quá trình lấy mẫu ngẫu nhiên sẽ dẫn đến dữ liệu $D = \{x(1), \dots, x(n)\}$, trong đó $x(i)$ là vector p chiều thứ i của độ đo, tức dòng thứ i trong ma trận dữ liệu $n \times p$. Xác suất này được thể hiện qua $p(D | M)$. Thông thường người ta không đặt tường minh phụ thuộc trên mô hình M , mà viết đơn giản là $p(D)$, phù hợp với ngữ cảnh.

Cho $p(x(i))$ là xác suất của cá nhân i có độ đo vector $x(i)$; ở đây p là hàm xác suất hoặc hàm mật độ tùy theo bản chất của x . Hơn nữa, giả sử xác suất của mỗi phần tử của dân cư được chọn đưa vào mẫu không ảnh hưởng đến xác suất của phần tử đã được chọn, tức các quan sát tách biệt là độc lập, tức dữ liệu được rút ngẫu nhiên, thì xác suất toàn thể của quan sát phân phối toàn thể của các giá trị trong mẫu sẽ đơn giản là tích các xác suất thành phần $p(D | \theta, M) = \prod_{i=1}^n p(X(i) | \theta, M)$; trong đó M là mô hình và θ là tham số của mô hình. Khi xét đến tham số này của mô hình, M

¹ Model: mô hình; Pattern: khuôn mẫu.

được gọi là *hàm phù hợp*¹. Các phương pháp được phát triển ứng với tình huống này, trong đó việc quan sát một giá trị sẽ làm thay đổi cơ hội quan sát giá trị khác, nhưng sự độc lập vẫn có.

Dựa trên xác suất này, người ta có thể xác định mức hiện thực của mô hình giả định. Nếu nhờ tính toán người ta thấy mô hình ít có thể xảy ra, có thể loại bỏ mô hình. Mô hình sẽ bị loại nếu dữ liệu quan sát được theo mô hình đó thấp hơn ngưỡng xác định, thường là 0.01 hay 0.05, tức ngưỡng có ý nghĩa của việc thử.

Nguyên tắc tương tự được dùng trong đánh giá các giá trị dân cư đối với các tham số của mô hình. Giả sử mô hình cho thấy dữ liệu hiện theo phân phối chuẩn với sai khác đơn vị với trung bình chưa biết μ . Người ta đề xuất các giá trị trung bình khác nhau, rồi tính xác suất để dữ liệu quan sát hiện ra nếu trung bình dân cư có giá trị đó. Người ta thử nghiệm, và bỏ các giá trị có xác suất thấp đi. Hoặc người ta có thể rút ngắn quá trình bằng cách đánh giá trung bình ứng với xác suất cao nhất do dữ liệu quan sát tạo nên. Giá trị này được gọi là đánh giá thuận cực đại về trung bình, và quá trình là đánh giá thuận cực đại. Khi thể hiện xác suất để mô hình cụ thể có được do dữ liệu quan sát, như hàm có tham số, được gọi là hàm thuận. Dùng hàm này để xác định khoảng giá trị hợp lý.

5.5.5. Ước lượng

Đã có một số kĩ thuật tóm tắt dữ liệu trong phần trước. Tiếp cận suy diễn, cần có các khẳng định tổng quát hơn về toàn thể dân cư. Đó là khẳng định về phân phối xác suất hay hàm mật độ xác suất mà các dữ liệu lộ ra, hay tương đương là hàm phân phối tích lũy.

5.5.5.1. Các tính chất ước lượng mong muốn

Tiếp theo là hai phương pháp quan trọng để ước lượng các tham số mô hình (i) ước lượng thuận cực đại; (ii) ước lượng Bayes.

¹ Likelihood function: hàm phù hợp.

Các phương pháp khác nhau sẽ có tính chất khác nhau, tùy theo bài toán.

Cho $\bar{\theta}$ là cái ước lượng của tham số θ . Do $\bar{\theta}$ là con số rút ra từ dữ liệu; nó thay đổi với dữ liệu khác nhau. Do vậy nó có phân phối, với giá trị khác nhau ứng với thay đổi của lấy mẫu. Người ta có thể có mô tả tóm tắt của phân phối. Cũng có thể dùng trung bình hay giá trị kì vọng $E[\bar{\theta}]$. Hàm E hợp với phân phối đúng, tuy chưa biết, với dữ liệu lấy mẫu, tức trên tất cả dữ liệu có thể được, kích thước n với trọng số là xác suất xuất hiện.

Độ lệch của $\bar{\theta}$ được xác định nhờ hàm *lệch* $(\theta) = E[\bar{\theta}] - \theta$, là độ lệch giữa giá trị kì vọng của ước lượng $E[\bar{\theta}]$ và giá trị thực của tham số θ . Ước lượng không lệch khi $E[\bar{\theta}] = \theta$. Cái ước lượng này cho thấy không xuất phát một cách hệ thống từ các giá trị tham số thực trên trung bình, dù rằng đối với tập D cụ thể, thu được $\bar{\theta}$ khác với θ . Lưu ý phân phối mẫu và giá trị thực của θ là chưa biết, nên không thể tính được độ lệch đối với bộ dữ liệu đã biết.

Phương sai của độ lệch $\text{var}(\bar{\theta}) = E[\bar{\theta} - E[\bar{\theta}]]^2$. Phương sai do thành phần sai sót ngẫu nhiên dựa theo dữ liệu trong thủ tục ước lượng; thể hiện ước lượng nhạy cảm ra sao với bộ dữ liệu riêng. Phương sai không phụ thuộc vào giá trị thực của θ ; nó đo ước lượng thay đổi quanh dữ liệu quan sát ra sao. Do vậy cho dù chưa biết phân phối dữ liệu thực, có thể ước lượng được phương sai dựa trên dữ liệu của ước lượng, khi cho n , bằng cách tính lặp các phương sai của $\bar{\theta}$ trên các mẫu lấy nhiều lần. Ước lượng không chệch có giá trị phương sai nhỏ nhất, được gọi là *ước lượng không chệch tốt nhất*¹.

Sai số bình phương trung bình của $\bar{\theta}$ là $E[(\bar{\theta} - \theta)^2]$, trung bình của độ lệch bình phương giữa giá trị ước lượng và giá trị thực của tham số. Sai số bình phương trung bình có phân rã tự nhiên theo tổng độ lệch bình phương của $\bar{\theta}$ và phương sai của nó:

¹ The best unbiased estimator: ước lượng không chệch tốt nhất.

$$\begin{aligned} E[(\bar{\theta} - \theta)^2] &= E[(\bar{\theta} - E[\bar{\theta}] + E[\bar{\theta}] - \theta)^2] \\ &= (E[(\bar{\theta} - \theta)^2] + E[(\bar{\theta} - E[\bar{\theta}])^2]) = (\text{lech}(\bar{\theta}))^2 + \text{var}(\bar{\theta}) \end{aligned}$$

trong đó từ dòng 1 sang 2 có $E[\theta] = \theta$ do nó là hằng số.

Giả sử có dãy $\bar{\theta}_{n1}, \dots, \bar{\theta}_{nm}$ là các ước lượng dựa trên các mẫu có kích thước tăng $n_1, \dots, n_m$. Dãy được gọi là bền vững nếu xác suất sai khác giữa $\bar{\theta}$ và giá trị thực θ là lớn hơn giá trị bất kì, có khuynh hướng đến 0 khi tăng kích thước mẫu. Điều này tốt đối với khai phá dữ liệu, do mẫu càng lớn, ước lượng càng gần với giá trị thực.

5.5.5.2. Ước lượng thuận cực đại

Ước lượng thuận cực đại là phương pháp ước lượng có tham số dùng rộng rãi. Xét bộ dữ liệu $D = \{x(1), \dots, x(n)\}$, được lấy mẫu một cách độc lập từ cùng phân phối $f(x | \theta)$; tức phân phối độc lập và y hệt nhau như người làm thống kê nói. Hàm phù hợp $L(\theta | x(1), \dots, x(n))$ là xác suất để dữ liệu xuất hiện, với giá trị θ đã cho, được xem như hàm của θ , tức $p(D | \theta)$. Dù có vai trò của mô hình M , không nhất thiết liệt kê.

Do giả thiết các quan sát là độc lập, có

$$L(\theta | D) = L(\theta | x(1), \dots, x(n)) = p(x(1), \dots, x(n)) = \prod_{i=1}^n f(x(i) | \theta)$$

trong đó θ là hàm vô hướng; tự θ có thể là vector của các tham số chứ không hẳn là tham số đơn. Sự phù hợp của bộ dữ liệu $L(\theta | D)$, xác suất của dữ liệu D quan sát được đối với mô hình cụ thể, là khái niệm cơ bản trong phân tích dữ liệu. Xác định sự phù hợp cho các vấn đề đặt ra để xác định mô hình xác suất về cách tạo ra dữ liệu. Xác định sự phù hợp cho phép các ứng dụng dùng suy diễn thống kê. Lưu ý do sự phù hợp được xác định như hàm của θ , có thể viết gọn $p(D | \theta)$ không có θ , tức sự phù hợp chỉ xác định với hằng vô hướng tùy ý, nên đáng hàm của θ không thể hiện giá trị thực mà nó nhận.

Giá trị đối với θ để dữ liệu có khả năng xuất hiện cao là ước lượng sự phù hợp cực đại, MLE¹. Người ta kí hiệu nó MLE đối với θ là $\bar{\theta}_{ML}$.

Ước lượng sự phù hợp cực đại là tốt theo trực giác hoặc theo toán học. Nếu $\bar{\theta}_{ML}$ là MLE với tham số θ , thì $g(\bar{\theta}_{ML})$ là MLE của hàm $g(\theta)$, dù để cẩn thận cần xét g là hàm một một. Mặt khác không thể cầu toàn; ước lượng sự phù hợp cực đại cũng bị lệch, tùy theo tham số và mô hình, dù lệch rất nhỏ đối với bộ dữ liệu lớn, thường là $O(1/n)$.

Đối với bài toán đơn giản, tức cấu trúc toán học đơn giản chứ không phải dữ liệu, có thể tìm được MLE nhờ phép toán sai phân. Thực tế $L(\theta)$ sự phù hợp theo logarit là cực đại thường dùng, vì người ta thay tích bằng tổng, cho kết quả như $L(\theta)$ và hàm logarit là đơn điệu. Tuy nhiên người ta quan tâm đến mô hình hơn là một tham số. Định nghĩa sự phù hợp một biến được tổng quát hóa trực tiếp cho trường hợp nhiều biến, nhưng hàm là nhiều biến của d tham số, tức hàm vô hướng xác định trên không gian tham số d chiều. Do d có thể lớn, việc phát hiện cực đại của hàm d chiều không dễ.

Việc ước lượng điểm như trên là tốt, không gán với yếu tố không chắc chắn, do chấp nhận ước lượng gần tốt. Có thể dùng ước lượng khoảng, sử dụng khoảng tin cậy; cận trên, dưới của khoảng được gọi là các cận tin cậy. Về khoảng: nó được tính nhờ thủ tục, sẽ chứa θ với một giá trị xác suất; khoảng được tính từ mẫu và là biến ngẫu nhiên.

Khoảng tin cậy thường xuyên được dựa trên giả thiết rằng thống kê mẫu có phân phối chuẩn. Điều này là hiện thực: định lý giới hạn trung tâm² cho biết phân phối của nhiều thống kê có thể xấp xỉ tốt nhờ phân phối chuẩn, đặc biệt nếu kích thước mẫu lớn. Dùng xấp xỉ này, người ta tìm thấy khoảng trong đó biết được xác

¹ Maximum likelihood estimator: ước lượng khả năng cực đại.

² Central limit theorem: định lý giới hạn trung tâm.

suất về giá trị thực chưa biết xảy ra của tham số θ , và tìm thấy tiếp một khoảng cho tham số chưa biết. Để theo tiếp cận này, cần có ước lượng của độ lệch chuẩn của ước lượng $\bar{\theta}$. Một cách suy ra ước lượng đó là phương pháp *tự cố*¹.

5.5.5.3. Ước lượng Bayes

Tiếp cận thường xuyên về suy luận sử dụng các tham số về dân cư; tham số này cố định nhưng chưa rõ, và dữ liệu là các mẫu ngẫu nhiên từ dân cư. Các biến thực chất nằm trong $D = \{x(1), \dots, x(n)\}$.

Tiếp cận Bayes xem dữ liệu như đã biết, được quan sát và ghi lại. Tham số θ là biến ngẫu nhiên. Trong khi tiếp cận thường xuyên coi θ cố định và chưa biết, tiếp cận Bayes có phân phối với các giá trị sẽ nhận, và quan sát sẽ cho phép hiện các giá trị theo phân phối. Giá trị $p(\theta)$ thể hiện mức độ tin rằng tham số θ là đúng. Nếu hàm này có chóp đỉnh, người ta tin hơn so với hàm có dạng phẳng.

Tiếp cận Bayes được dùng nhiều trong công nghệ thông tin, trong mô hình xác suất của phân tích dữ liệu.

Trước khi phân tích dữ liệu, phân phối xác suất để θ nhận các giá trị khác nhau là phân phối *tiền nghiệm*² $p(\theta)$. Việc phân tích dữ liệu D sẽ làm thay đổi phân phối, thu được phân phối *hậu nghiệm*³ $p(\theta | D)$. Thay đổi từ trước sang sau tuân theo định lý mang tên Thomas Bayes:

$$p(\theta | D) = \frac{p(D | \theta) \cdot p(\theta)}{p(D)} = \frac{p(D | \theta) \cdot p(\theta)}{\int_{\psi} p(D | \psi) \cdot p(\psi) d\psi}$$

Thủ tục thay đổi này cho phép thu được phân phối chứ không phải giá trị của θ . Từ phân phối, người ta ước lượng giá trị đơn. Người ta có thể xem xét nghĩa của phân phối hậu nghiệm và kiểu của phân phối; kỹ thuật này được gọi là *phương pháp cực đại*

¹ Bootstrap: tự cố.

² Prior: tiền nghiệm.

³ Posterior: hậu nghiệm.

*hậu nghiệm*¹ MAP. Nếu chọn phân phối tiên nghiệm $p(\theta)$ theo cách đặc biệt, thí dụ $p(\theta)$ đồng nhất, MAP và ước lượng thuận cực đại về θ có thể sát nhau, do phân phối tiên nghiệm phẳng, không chọn được θ nào. Theo cách này, thuận cực đại được coi như trường hợp đặc biệt của thủ tục MAP; thủ tục này là dạng hạn chế, ước lượng điểm, của ước lượng Bayes.

Đối với bộ dữ liệu đã biết D và mô hình cụ thể, mẫu số của công thức trên là hằng, nên có thể viết lại $p(\theta | D) \propto p(D | \theta) \cdot p(\theta)$. Phân phối hậu nghiệm của θ với D đã biết, tức phân phối với điều kiện đã có các quan sát dữ liệu D , tỉ lệ với tích của xác suất trước $p(\theta)$ và khả năng $p(D | \theta)$. Nếu chỉ tin ít về giá trị của các tham số trước khi thu thập dữ liệu, người ta sẽ chọn trước để phổ xác suất rộng. Trong trường hợp này, bộ dữ liệu quan sát càng lớn, sự phù hợp càng thể hiện trong phân phối hậu nghiệm, và độ quan trọng của dạng cụ thể của xác suất trước càng thấp.

Một trong các tính chất nổi bật của tiếp cận Bayes là tránh được ước lượng điểm, như ước lượng sự phù hợp của tham số, khi thu thập tri thức của điều không chắc chắn trong bài toán, chẳng hạn tính phân phối hậu nghiệm toàn bộ về θ .

Thí dụ sử dụng tiếp cận Bayes để dự đoán về điểm dữ liệu mới $x(n+1)$, là điểm dữ liệu không thuộc bộ dữ liệu huấn luyện D . Thay vì sử dụng điểm ước lượng cho $\bar{\theta}$ trong mô hình dự đoán, tiếp cận Bayes lấy trung bình các giá trị θ có gắn trọng số xác suất sau $p(x(n+1) | D) = \int p(x(n+1), \theta | D) d\theta = \int p(x(n+1) | \theta) \cdot p(\theta | D) d\theta$, do $x(n+1)$ độc lập có điều kiện với bộ dữ liệu D với θ đã cho. Thực tế người ta cũng lấy trung bình theo các mô hình khác nhau, với kỹ thuật gọi là trung bình mô hình Bayes. Về bản chất, tất cả các trung bình này có thể đúng hơn về tính toán so với tiếp cận sự phù hợp cực đại. Đây là lí do trước tiên để các phương pháp Bayes có giá trị thực tế trong vài năm lại đây, ít ra đối với các bộ dữ liệu

¹ Maximum a posterior method: phương pháp cực đại hậu nghiệm.

quy mô nhỏ. Đối với bộ dữ liệu lớn và số chiều cao, phương pháp phân tích Bayes đầy đủ có thể tốt.

Việc chọn phân phối tiên nghiệm có thể quan trọng đối với phân tích Bayes, quan trọng hơn đối với các mẫu nhỏ. Phân phối tiên nghiệm thể hiện mức độ tin cậy ban đầu để tham số nhận giá trị khác nhau. Người ta càng tin cậy thì xác suất trước càng sát với giá trị; tin cậy càng ít thì càng tán xạ. Trong trường hợp theo trung bình chuẩn, khi không có ý niệm về giá trị thực, người ta muốn sử dụng phân phối tiên nghiệm đối với mỗi giá trị, tức phân phối tiên nghiệm là hoàn toàn phẳng hay biến thiên vô hạn. Điều này không ứng với bất kì hàm mật độ *đúng*¹, là hàm có vài giá trị 0 và tiến đến 1. Tuy nhiên đôi khi vẫn chấp nhận phân phối tiên nghiệm không thật đúng, nhưng nhất quán, đồng dạng trong toàn bộ không gian của các tham số. Có thể xem phân phối tiên nghiệm đó là phẳng trong tất cả các vùng mà tham số có mặt. Cho dù chấp nhận như thế, các phân phối tiên nghiệm đã đồng dạng đối với một tham số cụ thể lại không đồng dạng đối với việc chuyển hóa phi tuyến của tham số này.

Một điều liên quan đến suy luận Bayes, phân phối tiên nghiệm cho biết tin cậy trước của cá nhân vào giá trị có thể của tham số, và các cá nhân không giống nhau. Điều này phù hợp trường hợp này, nhưng với trường hợp khác lại không. Người ta khắc phục bằng cách sử dụng *thỏa thuận*² suy luận phù hợp. Dạng thỏa thuận suy luận thông dụng là của Jeffrey.

$$\text{Trước hết xác định thông tin Fisher } I(\theta|x) = -E\left[\frac{\partial^2 \log L(\theta|x)}{\partial \theta^2}\right],$$

đối với tham số vô hướng θ , tức kì vọng của đạo hàm bậc 2 của sự phù hợp log. Về cơ bản nó cho phép đo độ phẳng hay cong của hàm phù hợp. Hàm sự phù hợp càng phẳng, thông tin cung cấp về

¹ Proper: đúng.

² Prior: thỏa thuận.

giá trị tham số càng ít. Thỏa thuận của Jeffrey được xác định là $p(\theta) \propto \sqrt{I(\theta|x)}$.

Đây là thỏa thuận thích hợp do nếu $\phi = \phi(\theta)$ là hàm của θ , có thỏa thuận tỉ lệ với $\sqrt{I(\phi|x)}$. Điều này có nghĩa thỏa thuận bền vững sẽ có tác dụng cho dù tham số được biến dạng hay không.

Đã có nhận xét rằng việc ước lượng điểm đơn từ phân phối hậu nghiệm là đơn giản. Ước lượng khoảng cũng dễ, bằng cách tích hợp phân phối hậu nghiệm trên miền để biết xác suất tham số thuộc miền. Khi cần tham số đơn và miền là một khoảng, kết quả là *khoảng tin cậy*¹. Khoảng tin cậy là khoảng chứa xác suất đã cho, chừng 90%, để mật độ sau là cao nhất trong khoảng. Khi chấp nhận kí pháp Bayes với tham số là biến ngẫu nhiên, việc diễn giải khoảng này là dễ hơn so với diễn giải với tiếp cận các khoảng tin cậy thường xuyên.

Tuy nhiên, hiếm có mô hình chỉ cần đến một tham số. Mô hình điển hình cần vài tham số. Khi đó, người ta có thể tìm thấy phân phối *hậu nghiệm đồng thời*² đối với tất cả các tham số một cách đồng thời hoặc đối với chỉ tập cụ thể các tham số. Cũng có thể nghiên cứu các phân phối điều kiện đối với các tham số đã có giá trị cố định. Cho đến nay, thống kê Bayes đảm bảo suy luận và quy nạp, nhưng ít giá trị thực tế. Việc phân tích lẻ của các tham số riêng, từ các phân phối liên kết phức tạp là công việc khó khăn. Các năm qua cũng có các kết quả về phương pháp ước lượng nhằm ngẫu nhiên, dựa trên các mẫu ngẫu nhiên trong phân phối được ước lượng, dựa trên các thuộc tính của phân phối của các tham số. Các phương pháp này được gọi là *xích Markov Monte Carlo* MCMC³.

Các đặc tính đầu tiên của thống kê Bayes là ở xử lí tính chắc chắn. Người ta làm hiện các điều không chắc chắn trong phân tích dữ liệu, gồm cả không chắc chắn về các tham số được ước lượng,

¹ Credibility interval: khoảng tin cậy.

² Joint posterior distribution: phân phối hậu nghiệm đồng thời.

³ Markov Chain Monte Carlo: xích Markov Monto Carlo.

lần không chắc chắn về mô hình. Theo tiếp cận phù hợp, ước lượng điểm của tham số thường được coi là đích chính, nhưng phân tích Bayes liệt kê các phân phối hậu nghiệm của tham số, và cả phân phối hậu nghiệm của các cấu trúc mô hình. Dự đoán Bayes dùng trung bình trọng số theo giá trị tham số và cấu trúc mô hình. Các trọng số tỉ lệ với phù hợp của tham số, mô hình với dữ liệu, thời gian, phân phối tiên nghiệm đã cho. Về nguyên tắc, trung bình trọng số này có thể chính xác hơn tiếp cận các *phương án thay thế*¹ về điều kiện trên mô hình đơn với ước lượng điểm tham số. Tuy nhiên thực tế tiếp cận Bayes đòi hỏi ước lượng trọng số trung bình; khó khăn khi số chiều bài toán là lớn. Hơn nữa, trung bình có trọng số của tham số hay mô hình là ít thích hợp để diễn giải nếu việc mô tả là đích chính.

5.5.6. Kiểm định giả thuyết thống kê

Dù rằng khai phá dữ liệu liên quan đến việc tìm khía cạnh không nghi ngờ trong dữ liệu, khác với kiểm tra giả thuyết đặc biệt đưa ra trước khi thấy dữ liệu. Trong thực tế, người ta muốn kiểm tra giả thuyết.

Trong nhiều trường hợp, người ta muốn xem dữ liệu có chứng tỏ các ý tưởng về giá trị tham số. Do không thể đo toàn bộ dân cư, cần sử dụng mẫu. Các công cụ thống kê khảo sát các giả thuyết được gọi là *thử giả thuyết*².

5.5.6.1. Thử giả thuyết cổ điển

Nguyên tắc cơ bản của thử giả thuyết: bắt đầu xác định hai giả thuyết phụ, là giả thuyết rỗng và giả thuyết thay thế. Đôi khi giả thuyết rỗng là giá trị điểm, tức tác động với câu hỏi là không, tức không có sự khác nhau hay nghiêng theo hồi quy. Giả thuyết thay thế là phần bù của giả thuyết rỗng.

¹ Alternative approach: tiếp cận thay thế.

² Hypothesis test: thử nghiệm giả thuyết.

Chẳng hạn cần có kết luận về tham số θ . Giả thuyết rỗng, kí hiệu H_0 lấy $\theta = \theta_0$, giả thuyết thay thế H_1 lấy $\theta \neq \theta_0$. Sử dụng dữ liệu quan sát, tính được thống kê phù hợp nhất với bản chất của giả thuyết được thử. Nhà thống kê thay mẫu này, mẫu khác, như biến ngẫu nhiên. Nếu giả thuyết rỗng là chính xác, người ta có thể xác định phân phối mong muốn đối với thống kê đã chọn, và giá trị quan sát của thống kê sẽ là một điểm từ phân phối đó. Nếu các giá trị quan sát ra phía rìa của phân phối, người ta kết luận hoặc xảy ra sự kiện không hợp lí, hoặc giả thuyết rỗng là không thực tế. Giá trị quan sát càng đặc thù, người ta càng ít tin cậy trong giả thuyết rỗng.

Có thể đưa các con số vào thủ tục này. Xét đuôi trên của phân phối thống kê, với phân phối dựa trên giả thiết về giả thuyết rỗng, người ta có thể tìm thấy các giá trị tiềm năng có xác suất xảy ra khoảng 0.05. Đó là các giá trị đầu cực của thống kê. Các giá trị rút ra từ tập hợp các giá trị, với giả thiết giả thuyết rỗng là đúng. Nếu giá trị cực quan sát nằm trong miền trên, người ta có thể bỏ giả thuyết rỗng tại mức 5%: chỉ 5% thời gian hi vọng thấy kết quả tại vùng này nếu giả thuyết rỗng là đúng. Rõ ràng là vùng này được gọi là miền bỏ đi hay miền cực đoan. Tất nhiên, người ta không chỉ quan tâm sự lệch do giả thuyết rỗng theo một hướng. Có thể xét đuôi thấp hơn hay cao hơn trong phân phối. Trong trường hợp này có thể xác định miền bỏ đi như hợp của giá trị thống kê thử trong phần trên 2.5% của phân phối xác suất và giá trị thống kê thử trong phần dưới 2.5%. Đây là thử hai đuôi, khác với thử một đuôi vừa nêu. Kích thước của miền bỏ đi, được xem như mức có ý nghĩa của thử có thể được chọn tùy ý. Giá trị thông thường là 1%, 5% và 10%. Người ta có thể so sánh các thủ tục thử khác nhau theo *độ mạnh*¹ của nó. Độ mạnh của việc thử là xác suất để việc loại bỏ giả thuyết rỗng sai. Để ước lượng độ mạnh của việc thử, người ta cần giả thuyết thay thế để có thể tính xác suất để thống kê thử rơi vào vùng loại bỏ khi giả thuyết thay thế là đúng. Câu hỏi cơ bản là cách phát hiện thống kê thử tốt đối với bài toán cụ thể. Một chiến

¹ Power: độ mạnh.

lược là sử dụng *tỉ số phù hợp*¹. Thống kê tỉ số phù hợp được dùng để thử giả thuyết $H_0: \theta = \theta_0$ trái với $H_1: \theta \neq \theta_0$, được xác định

$$\lambda = \frac{L(\theta_0 | D)}{\sup_{\psi} L(\psi | D)}, \text{ trong đó } D = \{x(1), \dots, x(n)\}. \text{ Điều này có nghĩa}$$

tỉ số của sự phù hợp khi $\theta = \theta_0$ là lớn nhất về sự phù hợp khi θ không bị ràng buộc. Rõ ràng là giả thuyết rỗng bị loại khi λ nhỏ. Thủ tục được tổng quát hóa cho trường hợp giả thuyết rỗng không là giả thuyết một điểm mà có tập các giá trị đối với θ .

Có loại thử được dùng thường xuyên. Đó là thử về sai khác giữa trung bình, thử để so sánh phương sai, và thử để so sánh một phân phối quan sát với phân phối giả thuyết, được gọi là *kiểm định sự phù hợp*².

Chiến lược thử giả thuyết nêu trên dựa trên giả thiết là mẫu ngẫu nhiên được lấy ra từ phân phối nào đó, và đích của thử là khẳng định xác suất về tham số của phân phối đó. Đối tượng về sau là suy luận từ mẫu ra giá trị tiềm ẩn trong dân cư. Do vậy đôi khi người ta gọi nó là *lấy mẫu*³. Một chiến lược khác đôi khi phù hợp, đặc biệt khi người ta không tin rằng mẫu thu được từ lấy mẫu xác suất, và do vậy không thể suy luận với dân cư. Trong trường hợp này, người ta có thể vẫn có khẳng định xác suất về một vài tác động do giả thuyết rỗng. Chẳng hạn so sánh trị liệu và nhóm điều khiển. Người ta có thể chấp nhận giả thuyết rỗng với không có tác động trị liệu, nên phân phối của tỉ số những người nhận được trị liệu sẽ như của người không được trị liệu. Nếu lấy mẫu của những người, có thể không lấy ngẫu nhiên, và gán ngẫu nhiên họ với trị liệu và nhóm điều khiển, người ta hi vọng sai khác của tỉ số trung bình giữa các nhóm là nhỏ nếu giả thuyết rỗng là đúng. Vậy, theo giả thiết công bằng, khó ra khỏi phân phối về sai khác giữa trung bình mẫu của hai nhóm người ta mong đợi nếu không có tác động trị liệu, và nếu sai khác này là hệ quả của sự

¹ Likelihood ratio: tỉ số phù hợp.

² Goodness - of - fit test: kiểm định sự phù hợp.

³ Sampling paradigm: khung mẫu.

không cân bằng trong bố trí ngẫu nhiên. Người ta có thể khám phá sự vô lý về sai khác là lớn và lớn hơn cái thấy được. Các việc thử dựa trên nguyên tắc này có thuật ngữ *thử ngẫu nhiên* hay *thử giao hoán*¹. Lưu ý rằng không có suy luận thống kê từ mẫu cho toàn thể dân cư, nhưng có thể có khẳng định xác suất có điều kiện về tác động trị liệu, có điều kiện theo các giá trị quan sát.

Nhiều thử thống kê dùng giả thiết về dạng của phân phối dân cư được tiến hành lấy mẫu. Vậy nên đôi khi không phù hợp. Có lẽ người ta ít kiểm chứng giả thiết này, hay có lẽ người ta biết rằng dữ liệu không theo dạng của phép thử chuẩn. Trong hoàn cảnh này, người ta chấp nhận *thử phân phối tự do*². Các phép thử dựa trên hạng rơi vào lớp này. Tại đây, dữ liệu cơ bản được thay thế bằng nhãn đánh số theo vị trí chúng xảy ra. Chẳng hạn để khai thác hai mẫu có hiện ra trong cùng phân phối hay không, người ta sẽ thay thế các giá trị số hiện tại bằng hạng của chúng. Nếu chúng không hiện ra từ cùng một mẫu, người ta hi vọng các hạng của các phần tử của hai mẫu được trộn tốt. Tuy nhiên, nếu một phân phối có trung bình lớn hơn cái kia, người ta hi vọng một mẫu có khuynh hướng có hạng lớn và cái kia nhỏ. Nếu phân phối có cùng trung bình, nhưng một mẫu có phương sai lớn hơn, người ta sẽ hi vọng một mẫu cho thấy sự thừa các hạng lớn và nhỏ, và cái kia chiếm các hạng trung bình. Các thống kê thử có thể được cấu trúc dựa trên các giá trị trung bình hay số đo khác về hạng, và các mức ý nghĩa của chúng có thể được ước lượng theo đôi số ngẫu nhiên. Các thống kê thử như vậy gồm thống kê thử dấu hiệu, thống kê thử tổng hạng, thống kê thử Kolmogorov - Smirnov, và thống kê thử Wilcoxon. Đôi khi thuật ngữ thử không tham số được dùng, để mô tả các phép thử mà cơ bản là các phép thử không thử với giá trị của tham số của phân phối giả thiết nào.

So sánh giả thuyết H_0 và H_1 theo Bayes có được do so sánh các xác suất sau của chúng $p(H, | x) \propto p(x | H,) . p(H,)$.

¹ Randomization tests: thử ngẫu nhiên; Permutation tests: thử giao hoán.

² Distribution free test: thử tự do về phân phối.

$$\text{Từ đó có nhân tố Bayes}^1 \frac{p(H_0 | x)}{p(H_1 | x)} \propto \frac{p(H_0)}{p(H_1)} \frac{p(x | H_0)}{p(x | H_1)}.$$

Tuy nhiên cũng nảy sinh vài phức tạp. Có thể sự phù hợp lẽ có được do tích phân theo các tham số không được xác định trong giả thuyết, và xác suất trước sẽ là 0 nếu H_i hướng đến giá trị cụ thể của các giá trị. Tức là nếu chúng hướng đến giá trị của tham số θ , trong đó θ có thể lấy giá trị giữa 0 và 1. Một chiến lược xử lý việc này là gán xác suất khác 0 cụ thể cho giá trị θ .

5.5.6.2. Thử giả thuyết theo ngữ cảnh

Phần này đã trình bày tiếp cận cổ điển, thường dùng, để thử giả thuyết thống kê. Trong khai phá dữ liệu, các phân tích càng phức tạp. Sau đây là vài lý do:

1. Khai phá dữ liệu cần khối lượng lớn các dữ liệu, người ta hi vọng thu được ý nghĩa về thống kê: cho dù ban đầu mô hình giả thuyết sát với điều có ý nghĩa, lúc thực hành chẳng thu được điều gì quan trọng. Nguy hại hơn, mô hình ứng với dữ liệu vận vẹo lại cho ý nghĩa. Đây là điều không tránh khỏi, nên để ý đến;
2. Thường thường mô hình khớp dần dần quá trình. Người ta làm tinh mô hình bằng cách thêm, bớt một số thuật ngữ. Khi chạy thử mỗi mô hình sẽ dẫn đến xác suất không đúng. Các thủ tục thử tuần tự được phát triển, nhưng phức tạp. Mặt khác, thử nhiều lần cũng khó khăn;
3. Việc coi khai phá dữ liệu là quá trình khám phá đã bị hiểu theo nhiều cách. Một trong các mô hình sẽ được xem xét.

Giả sử thử m giả thuyết rằng đúng, dù người ta không hiểu, với mức 5%; mỗi lần dựa trên bộ dữ liệu con, độc lập với lần thử khác. Đối với mỗi giả thuyết phân tách, có xác suất 0.05 loại bỏ không đúng về giả thuyết này. Do thử là độc lập, xác suất loại bỏ không đúng ít nhất một giả thuyết là $1 - (1 - 0.05)^m$. Khi $m = 1$,

¹ Bayes factor: nhân tố Bayes.

$p = 0.05$, là tốt. Khi $m = 10$, $p = 0.4013$; $m = 100$, $p = 0.0041$. Do vậy nếu người ta thử khoảng 100 giả thuyết rỗng, người ta sẽ loại bỏ ít ra là một giả thuyết. Người ta có thể điều khiển tỉ lệ sai, bằng cách thiết lập xác suất loại bỏ không đúng một hay nhiều trong số m giả thuyết rỗng là 0.05. Trong trường hợp này, $0.05 = 1 - (1 - \alpha)^m$; với $m = 10$, $\alpha = 0.0051$; $m = 100$, $\alpha = 0.0005$. Điều này có nghĩa người ta có xác suất nhỏ đối với loại bỏ không đúng bất kì thành phần tách biệt của các giả thuyết nào.

Trong thực tế, hiện tượng phức tạp hơn nhiều; các giả thuyết không hoàn toàn độc lập. Nếu mà chúng hoàn toàn phụ thuộc thì việc chấp nhận hay loại bỏ một kéo theo chấp nhận hay loại bỏ tất cả. Giả thuyết không độc lập với cấu trúc phụ thuộc chưa biết, và điển hình là trộn lẫn của giả thuyết rỗng đúng với giả thuyết rỗng sai.

Một loạt các thủ tục thử đồng thời đã được phát triển để khắc phục khó khăn trên. Đây là vấn đề cơ bản. Tiếp cận cơ bản dựa trên *bất đẳng thức Bonferroni*¹. Người ta phát triển xác suất $(1 - \alpha)^m$ để không giả thuyết rỗng nào bị loại bỏ. Do $1 - (1 - \alpha)^m = m\alpha$, xác suất để một hay nhiều giả thuyết rỗng đúng bị loại không đúng là nhỏ hơn hay bằng $m\alpha$. Nhìn chung, xác suất loại bỏ không đúng một hay nhiều giả thuyết rỗng đúng là nhỏ hơn tổng xác suất của loại bỏ không đúng mỗi giả thuyết. Đây là tính không bằng nhau Bonferroni² bậc 1. Có thể có ước lượng sát hơn, với tri thức về mối quan hệ phụ thuộc giữa các giả thuyết.

Với vài thủ tục thử, có thể khó khăn khi việc thử cả họ các giả thuyết bỏ đi giả thuyết rỗng, mà người ta tin rằng ít nhất có một cái sai, nhưng không thành phần đơn nào bị loại bỏ. Các chiến lược được phát triển dùng cho ứng dụng cụ thể. Chẳng hạn trong phân tích nhiều chiều về phương sai, trong đó so sánh vài nhóm các đối tượng được đo theo nhiều biến. Các thủ tục thử được phát triển theo cách so sánh mỗi thống kê thử với giá trị ngưỡng đơn.

¹ Bonferroni inequality: bất đẳng thức Bonferroni.

² Trong thống kê, hiệu chỉnh Bonferroni là một trong các phương pháp tránh vấn đề so sánh nhiều lần; Do nhà toán học Italy, Carlo Emilio Bonferroni đề xuất.

Qua trao đổi trên, rõ ràng là việc đưa xác suất khẳng định các loại khác nhau, qua các thử giả thuyết, có vị trí trong khai phá dữ liệu; nhưng chưa là giải pháp vạn năng. Tuy nhiên chúng được coi như loại cụ thể của thủ tục tổng quát, cho phép khớp dữ liệu và khẳng định với giá trị số hay đích. Đích cao cho thấy khẳng định hay mô hình được tham chiếu, mà không cần giải thích xác suất nào. Các hàm đích khít với ý muốn thuộc loại này.

5.5.7. Các phương pháp lấy mẫu

Đặc trưng của khai phá dữ liệu là phân tích thứ cấp; người khai phá dữ liệu không trực tiếp phải thu thập dữ liệu. Tuy nhiên nếu có thông tin về quá trình thu thập dữ liệu, việc khai phá dữ liệu cũng tốt hơn. Thu thập dữ liệu thống kê truyền thống thường là trả lời một số câu hỏi điều tra về tính hiệu quả. Do khai phá dữ liệu là tìm cái không mong đợi, hay cái chưa thấy trước, các câu hỏi đi sâu vào bộ dữ liệu đã thu thập được. Do vậy người ta cũng không đi sâu về thiết kế theo kinh nghiệm, trong đó tối ưu hóa việc thu thập dữ liệu. Thật ra, người khai phá dữ liệu không điều khiển bộ dữ liệu lại có thể giải thích được chất lượng dữ liệu kém: dữ liệu có thể theo mục đích thu thập nhưng không tốt đối với việc khai phá dữ liệu.

Cơ sở dữ liệu chứa toàn bộ dân cư, các kí pháp suy luận thống kê không thích hợp, do khi người ta muốn biết tham số dân cư nào đó, người ta có thể tính ra. Tất nhiên điều này giả thiết rằng dữ liệu hoàn toàn mô tả dân cư, không có sai sót về đo, về dữ liệu thiếu... Do điều này là không thể, vẫn cần suy luận trên dữ liệu. Ngoài ra, kí pháp của dân cư và các mẫu có thể bị hiểu sai. Chẳng hạn khi đã có các giá trị về dân cư trong cơ sở dữ liệu, thường đích không phải là mô tả bộ dữ liệu, mà là khẳng định về các giá trị tương lai. Điều này cần đến sự không chắc chắn. Đặc biệt người ta quan tâm đến dự đoán. Phần trước đã phân biệt hai cách lấy mẫu trong khai phá dữ liệu (i) vài cơ sở dữ liệu chỉ là một mẫu từ dân cư lớn hơn; (ii) cơ sở dữ liệu có các bản ghi đối với

mọi đối tượng trong dân cư, nhưng phân tích dữ liệu chỉ dựa trên một mẫu từ các dữ liệu. Kỹ thuật thứ hai phù hợp chỉ với mô hình hóa tình huống và phát hiện mẫu; nó không phù hợp đối với thao tác tìm bản ghi riêng.

Đích đặt ra là rút ra mẫu từ cơ sở dữ liệu để cho phép tạo nên mô hình, thể hiện cấu trúc dữ liệu trong cơ sở dữ liệu. Lí do của việc chỉ dùng mẫu, thay cho toàn bộ cơ sở dữ liệu, là hiệu quả. Việc dùng toàn bộ dữ liệu là không khả thi, theo nghĩa thời gian tính toán. Bằng cách tính toán trên mẫu, người ta lợi về thời gian. Dĩ nhiên, mẫu cần thể hiện đặc tính của cơ sở dữ liệu. Nếu tỉ lệ lấy mẫu là 0.5, chọn 1 trong 2 bản ghi là đủ. Tiếp cận này được dùng với thuật ngữ *lấy mẫu hệ thống*¹. Thường thì nó chấp nhận được. Tuy nhiên nó cũng dẫn đến vấn đề không ngờ được. Nhìn chung, trong bất kì lược đồ lấy mẫu, mà các trường hợp lấy ra từ các mẫu chính quy, người ta không tính hết được sự phù hợp xảy ra rồi ro về tương tác với tính chính quy, tức mẫu lựa chọn ngẫu nhiên.

Thuật ngữ *ngẫu nhiên* vừa nêu với nghĩa tránh chuẩn tắc, chính quy. Nó cho biết xác suất để bản ghi được đối với mẫu. Các mẫu ngẫu nhiên có thể được dùng làm cơ sở cho suy luận thống kê: người ta có thể khẳng định về sự hợp lí của sự khác nhau giữa trung bình mẫu và trung bình dân cư.

Nếu lấy mẫu theo quá trình ngẫu nhiên, mẫu sẽ thỏa mãn tính không chính quy, và cả chính quy theo nghĩa hợp lí. Để tránh kết luận nhập nhằng, người ta thiết kế cơ chế lựa chọn mẫu để mỗi bản ghi có cơ hội lựa chọn như nhau. Một mẫu với xác suất như nhau về chọn các phần tử trong tập dân cư được gọi là mẫu *epsem*. Dạng cơ bản nhất của lấy mẫu *epsem* là lấy mẫu ngẫu nhiên đơn giản, trong đó n bản ghi trong mẫu được chọn từ N bản ghi của cơ sở dữ liệu theo cách n bản ghi có xác suất chọn như nhau. Ước lượng trung bình dân cư từ mẫu ngẫu nhiên đơn giản đúng bằng trung bình mẫu.

¹ Systematic sampling: lấy mẫu hệ thống.

Người ta phân biệt việc lấy mẫu với thay thế và lấy mẫu không thay thế. Theo cách đầu, một bản ghi đã được chọn trong mẫu lại có thể được rút; nhưng theo cách sau một khi bản ghi được rút, không thể rút lần thứ hai. Trong khai phá dữ liệu, do kích thước mẫu là nhỏ so với kích thước dân cư, khác nhau giữa kết quả của hai thủ tục được bỏ qua.

Mẫu ngẫu nhiên đơn là dạng cơ bản nhất của thiết kế mẫu; người ta cũng phát triển các phương pháp lấy mẫu khác cho các thuộc tính mong muốn theo điều kiện cụ thể. Trong phần này chỉ trình bày hai lược đồ quan trọng.

Theo cách *lấy mẫu ngẫu nhiên phân tầng*¹, toàn thể dân cư được tách thành tập dân cư con không đều nhau hay phân tầng, và một mẫu được rút tách biệt từ mỗi tầng. Mẫu này thường là ngẫu nhiên, nhưng không nhất thiết. Theo thủ tục này, có vài ưu điểm. Nó cho phép khẳng định về mỗi tập con dân cư. Ưu điểm khác: nếu các tầng là đồng nhất tương đối theo các biến quan tâm, phương sai của ước lượng chung có thể nhỏ hơn so với cách mẫu ngẫu nhiên. Chẳng hạn có công ty thẻ tín dụng chia các phạm trù ra 26 loại: siêu thị, lữ hành, trạm xăng... Người ta muốn ước lượng giá trị trung bình của các giao dịch, có thể lấy mẫu ngẫu nhiên đơn về các giá trị giao dịch từ các bản ghi cơ sở dữ liệu, tính trung bình để có kết quả ước lượng. Tuy nhiên với thủ tục này, một vài giao dịch không được thể hiện trong mẫu, bị ước lượng sai. Người ta có thể điều khiển quá trình bằng cách ép mẫu phải có một số giao dịch quan tâm. Điều này là theo cách phân tầng, trong đó các kiểu giao dịch phân tầng. Chẳng hạn cho biết lí do bên trong tầng là đồng nhất, trong khi dữ liệu không đồng nhất. Nếu tất cả các tầng có mức tán xạ như nhau như toàn dân cư, phương pháp theo tầng chẳng còn là ưu điểm.

Nhìn chung, giả sử muốn ước lượng trung bình dân cư đối với vài biến, sử dụng phương pháp lấy mẫu phân tầng, với việc lấy mẫu ngẫu nhiên đơn tại mỗi tầng. Giả sử tầng thứ k có N_k

¹ Stratified random sampling: lấy mẫu ngẫu nhiên phân tầng.

phần tử, và n_k trong đó được chọn cho mẫu của tầng đó. Kí hiệu trung bình mẫu r trong tầng thứ k là $\bar{x}_k$, ước lượng trung bình của toàn dân cư là $\sum_k \frac{N_k \bar{x}_k}{N}$, trong đó N là kích thước tổng của dân cư.

Phương sai của ước lượng này là $\frac{1}{N^2} \sum_k N_k^2 \text{var}(\bar{x}_k)$, trong đó $\text{var}(\bar{x}_k)$ là phương sai của mẫu ngẫu nhiên đơn kích thước n_k của tầng k .

Dữ liệu thường có cấu trúc phân cấp. Việc sinh ra khung lấy mẫu đầy đủ và rút ra được mẫu ngẫu nhiên đơn là khó. Các tệp trên các máy tính khác nhau nên càng khó. Chẳng hạn trong máy tính, các tệp cho phép rút ra mẫu về máy tính; chọn được các tệp đối với máy tính quan tâm và chuyển đến giai đoạn lấy mẫu tiếp theo.

Các cụm thường có kích thước không bằng nhau. Nếu tất cả các đơn vị của mỗi cụm được chọn, tỉ lệ lấy mẫu con là 1, mỗi đơn vị có xác suất được chọn là n / K , trong đó n là số cụm được chọn trong số tổng các cụm K . Nếu không chọn tất cả các đơn vị, mỗi đơn vị có cùng xác suất trong mẫu. Mẫu là epcsem. Đây là thiết kế thông thường. Việc ước lượng phương sai của thống kê dựa trên thiết kế đó khó hơn trong trường hợp mô tả trên, do kích thước mẫu cũng là biến ngẫu nhiên. Nó phụ thuộc vào các cụm có trong mẫu. Ước lượng trung bình của biến là tỉ lệ của hai biến: tổng toàn bộ các đơn vị trong mẫu và tổng đơn vị trong mẫu. Kí hiệu kích thước mẫu ngẫu nhiên được chọn từ cụm k là n_k , tổng toàn bộ các

đơn vị được chọn từ cụm k là sự kiện, trung bình mẫu r là $\frac{\sum_k x_k}{\sum_k n_k}$.

Nếu kí hiệu tỉ lệ lấy mẫu toàn bộ là f , thông thường f nhỏ và có thể bỏ qua, phương sai của r là

$$\frac{1-f}{(\sum_k n_k)^2} \cdot \frac{a}{1-a} (\sum_k s_k^2 + \sum_k n_k^2 - 2r \sum_k s_k n_k).$$

5.5.8. Nhận xét về độ chắc chắn

Thế giới thực cho thấy không gì là chắc chắn. Trong khai phá dữ liệu, đích là phát hiện trên dữ liệu. Người ta muốn được tin tưởng khi có khẳng định đúng, nhưng lại hay thỏa mãn với khẳng định sai. Khi phân tích toàn bộ dân cư, sự không chắc chắn sẽ giảm. Khi làm việc với mẫu, kết luận rút ra thường được áp dụng cho toàn bộ dân cư. Công cụ cơ bản sử dụng là xác suất. Áp dụng các kết quả của xác suất cho phép đạt được ước lượng tốt nhất về các giá trị, cho dù dữ liệu không đủ và chỉ vài mẫu được đo. Ngoài ra, ứng dụng kết quả xác suất giúp lượng hóa mức tin cậy về kết luận.

Tiếp theo cần xem xét về các tham số xác suất, được coi là công cụ khai phá dữ liệu, từ mô hình tổng quát đến xác định mẫu.

5.6. Kết luận

Mô hình là khía cạnh quan trọng trong khai phá dữ liệu, sau khi người ta đã chuẩn bị dữ liệu. Mô hình tốt sẽ cho đầu ra hợp lý, trên các dữ liệu mới. Các dữ liệu mới không là các dữ liệu (i) huấn luyện; (ii) thử nghiệm; (iii) kiểm tra mô hình.

Có hai khái niệm liên quan đến mô hình, là (i) mô hình, đối với toàn thể dữ liệu; (ii) khuôn mẫu, chỉ áp dụng đối với tập con của dữ liệu. Mô hình hay khuôn mẫu đều có các cấu trúc khác nhau.

Liên quan đến mô hình có khái niệm (i) độ phức tạp của mô hình; (ii) tính phù hợp của mô hình. Trong số mô hình xử lý dữ liệu, người ta quan tâm đến các trường hợp xử lý (i) dữ liệu bình thường; (ii) dữ liệu bất thường. Hai loại xử lý dữ liệu này được trình bày như các thí dụ về phân tích dữ liệu, hay khai phá dữ liệu.

Cũng liên quan đến mô hình xử lý dữ liệu là vấn đề về dữ liệu không chắc chắn. Phân tích dữ liệu không chắc chắn, quan điểm về xử lý loại dữ liệu được trình bày cuối chương.

Khai phá luật kết hợp

Vấn đề là xử lí dữ liệu lớn không được các kĩ thuật truyền thống đáp ứng. Với khối lượng lớn các dữ liệu, người ta cần có kĩ thuật xử lí dữ liệu, phát hiện tri thức cho phù hợp. Thực ra, việc xử lí dữ liệu lớn được đặt ra từ lâu, từ năm 1960 với hệ quản trị tệp.

Việc xử lí dữ liệu lớn thuộc về nhiều ngành khoa học: thống kê, quản trị dữ liệu, cơ sở dữ liệu, nhận dạng, trí tuệ nhân tạo... Điều này cho thấy việc xét khai phá dữ liệu chỉ có thể tập trung vào một số chủ đề, thuộc một phần các lĩnh vực khoa học vừa nêu.

Chương này nhằm cung cấp nguyên tắc cơ bản của khai phá dữ liệu, về phương pháp và kĩ thuật khai phá dữ liệu. Các chương trước đã đề cập một số kiến thức nền, liên quan đến quá trình khai phá dữ liệu. Tuy chưa đề cập các vấn đề ứng dụng thực tế của khai phá dữ liệu, nội dung chương này cũng cung cấp khái niệm về ứng dụng. Có thể tham khảo tài liệu của Han, Kamber, 2000, và của Witten, Frank, 2000, để có thêm thí dụ minh họa. Việc xét các nguyên tắc lí thuyết có thể đặt trên ứng dụng cụ thể để dễ minh họa ý tưởng, khái niệm hơn.

Do nội dung chương đề cập nhiều về thống kê và công nghệ thông tin; nó khiến người dùng cập nhật kiến thức, do hai lĩnh vực đều phức tạp. Để hiểu rõ khai phá dữ liệu, cần hiểu cả *mô hình toán học* lẫn *thuật toán tính toán*. Các phần sẽ gắn các kí pháp, giả thiết của mô hình thống kê với thế giới thực đối với các phương pháp, thuật toán tính toán.

Phần đầu chương là các nguyên tắc cơ bản của mô hình và suy luận; rồi đến các khung hệ thống cho phép liên kết các mô hình với dữ liệu qua các phương pháp và thuật toán; cuối cùng là thể hiện ý tưởng trên vấn đề cụ thể, về phân lớp và hồi quy. Vậy sẽ có (i) phần cơ bản, liên quan đến độ đo, thể hiện dữ liệu, suy luận; (ii) các thành phần của khai phá dữ liệu, với các thuật toán, việc thể hiện mô hình, tối ưu; (iii) nhiệm vụ khai phá dữ liệu và các thuật toán, liên quan đến một số vấn đề cụ thể và thuật toán đi kèm.

6.1. Tổng quan về khai phá dữ liệu

6.1.1. Giới thiệu về khai phá dữ liệu

Việc số hóa dữ liệu, nhiều lĩnh vực số hóa đã tạo nên kho dữ liệu lớn; nhưng vấn đề là *thông tin cần thiết* trong kho dữ liệu ấy. Khai phá dữ liệu là việc trích rút thông tin có giá trị từ các dữ liệu của cơ sở dữ liệu. Người ta cần xác định chính xác thuật ngữ khai phá dữ liệu.

Định nghĩa: Khai phá dữ liệu là việc phân tích các dữ liệu quan sát, thường là lớn, để phát hiện mối quan hệ tin cậy và tổng hợp dữ liệu theo cách mới, nhằm hiểu và sử dụng dữ liệu hiệu quả hơn.

Mối quan hệ được xem xét dưới dạng (i) *mô hình*; (ii) *khuôn mẫu*, như đã trình bày trong chương trước. Dữ liệu quan sát thường được coi là rộng hơn dữ liệu chọn lọc, nhờ kinh nghiệm. Các dữ liệu quan sát thường ở dạng thô, theo nhiều cách nhìn đa dạng, tạo nên kho dữ liệu lớn.

Trong bản thân khối dữ liệu thu thập được đã có sẵn *mối quan hệ*. Mối quan hệ này thể hiện thế giới thực có nhiều sự liên quan; ngoài ra, các phương pháp lưu trữ làm che mất tính thể hiện của mối quan hệ. Việc phát hiện các mối quan hệ và cấu trúc mới trên các dữ liệu là mới. Người ta có thể đặt ra giả thuyết rồi kiểm tra tính mới. Tính mới còn tùy thuộc vào tri thức của người dùng; và cần đơn giản, để người dùng dễ hiểu, cho dù chỉ mới về định tính, chưa cân đo, đo đếm định lượng.

Khai phá dữ liệu thường được xem là một phần của phát hiện tri thức. Việc khai phá tri thức là lĩnh vực thuộc trí tuệ nhân tạo.

Quá trình tìm mối quan hệ trong bộ dữ liệu, hay tìm thể hiện chính xác, phù hợp, có ích của khía cạnh nào đó của dữ liệu, cần có các bước:

1. Xác định *bản chất và cấu trúc* của thể hiện;
2. Xác định *cách định lượng* và *cách so sánh* các thể hiện khác nhau đối với dữ liệu; có nghĩa đưa ra được hàm đích;
3. Chọn quá trình *thủ tục* để tối ưu hàm đích;
4. Xác định *nguyên tắc* mà quản trị dữ liệu cần có, để thực hiện *thuật toán* một cách hiệu quả.

Khai phá dữ liệu là việc liên chuyên ngành, liên quan chính đến (i) thống kê; (ii) công nghệ cơ sở dữ liệu; (iii) học máy; (iv) trí tuệ nhân tạo; (iv) thể hiện đồ họa. Do vậy, khó xác định biên của chuyên ngành khai phá dữ liệu.

6.1.2. Bản chất của bộ dữ liệu

Định nghĩa: Bộ dữ liệu là tập các số đo được từ môi trường hay quá trình.

Đối với mỗi đối tượng, có p số đo; khi có n đối tượng, người ta có $n \times p$ số đo, tạo nên ma trận dữ liệu n dòng ứng với n đối tượng. Mỗi dòng có thể được coi như cá nhân, thực thể, trường hợp, đối tượng, bản ghi. Khái niệm này đã thấy trong mô hình dữ liệu quan hệ. Việc xác định các đối tượng, thực thể cần được chi tiết hóa.

Các cột của bảng được coi là biến, khía cạnh, thuộc tính, trường dữ liệu. Việc xác định tên gọi tùy theo ngôn ngữ trong lĩnh vực ứng dụng. Về các dữ liệu, có nhận xét sau:

- Dữ liệu *định lượng*, cho phép đo được giá trị;
- Một *thuộc tính* có thể quan trọng đối với đối tượng này, nhưng không quan trọng đối với đối tượng khác;

- Kỹ thuật phân tích dữ liệu có thể *phù hợp* với kiểu dữ liệu này, nhưng không phù hợp với kiểu dữ liệu của thuộc tính khác; chẳng hạn việc tính toán trên giá trị tuổi khác với tính toán trên số thứ tự của đối tượng;

Khung phân lớp dữ liệu ảnh hưởng đến việc đo dữ liệu. Theo khung phân lớp, dữ liệu có cấu trúc, có mối quan hệ đặc trưng. Dữ liệu có thể nhận các giá trị *liên tục* trong một khoảng, hay chỉ thường nhận một số giá trị xác định, tức *rời rạc*.

6.1.3. Các loại cấu trúc theo mô hình và khuôn mẫu

Các loại thể hiện thấy được trong khai phá dữ liệu được nhìn theo nhiều cách. Một trong các cách là đặc trưng qua mô hình tổng thể và khuôn mẫu cụ thể.

Định nghĩa: Cấu trúc mô hình là nhận xét tổng thể về bộ dữ liệu.

Cấu trúc mô hình cho phép thể hiện một điểm trong không gian đo bằng lời. Nếu coi dòng dữ liệu như vector trong không gian p chiều, mô hình cho phép phát biểu về một điểm trong không gian này; cho biết đã có giá trị hay chưa, và giá trị đó có thể được nội, ngoại suy. Chẳng hạn về mô hình đơn giản: quan hệ $Y = aX^2 + bX + c$; mô hình cho biết sự tuyến tính, theo thuật ngữ thống kê cổ điển, tuy dạng quan hệ hàm là phi tuyến.

Định nghĩa: Cấu trúc khuôn mẫu là nhận xét về miền hạn chế của không gian, theo một số biến

Xét thí dụ khẳng định xác suất IF $x > 1$ THEN $p(Y > 2) = p_1$. Cấu trúc này mang ràng buộc về biến X , Y , liên quan đến dạng luật xác suất.

Trái với cấu trúc tổng thể, các khuôn mẫu mô tả cấu trúc liên quan đến miền dữ liệu tương đối nhỏ trong không gian dữ liệu. Đôi khi cấu trúc mô hình và cấu trúc như là hai mặt đối lập của cùng vấn đề. Để phát hiện hành vi bất thường, người ta cần xét các hành vi bình thường, tổng thể. Ở đây có vai trò của chẩn đoán

trong phân tích thống kê; phương pháp phát hiện khuôn mẫu cho phép tìm hòng học, lỗi sai.

Trong các cấu trúc mô hình hay khuôn mẫu nêu trên, có các tham số; tức a, b, c đối với mô hình; $1, 2, p_1$ đối với khuôn mẫu. Nhìn chung, mỗi khi thiết lập dạng cấu trúc, người ta lưu ý việc phát hiện tham số, rồi đánh giá tham số trên các dữ liệu thực. Khi có giá trị tham số, người ta thu được mô hình cụ thể, chẳng hạn $Y = 2X^2 + 3X + 4$, như là *mô hình phù hợp*, hay *mô hình*. Với khuôn mẫu cũng tương tự. Việc phân biệt các *cấu trúc mô hình* với *mô hình* là quan trọng.

- Các cấu trúc thể hiện dạng tổng quát của các mô hình, hay khuôn mẫu, với các tham số chưa được xác định giá trị;
- Mô hình phù hợp có các giá trị riêng ứng với các tham số.

Việc phân biệt mô hình với khuôn mẫu có tác dụng trong nhiều trường hợp. Tuy nhiên theo thói quen đơn giản hóa, người ta không tách bạch mô hình và khuôn mẫu.

6.1.4. Các nhiệm vụ khai phá dữ liệu

Có thể chia công việc khai phá dữ liệu ra các nhiệm vụ. Nhiệm vụ ứng với chủ đề quan tâm của người phân tích dữ liệu. Dưới đây là một số nhiệm vụ; tuy nhiên có thể chia nhiệm vụ ra nhiệm vụ con, tùy theo loại hoạt động khai phá dữ liệu, và thuật toán tương xứng. Ở đây đưa ra năm nhiệm vụ.

6.1.4.1. Phân tích thăm dò dữ liệu, EDA¹

Đích của nhiệm vụ này là thăm dò dữ liệu mà chưa có ý tưởng hay định hướng nào. Các kĩ thuật EDA là tương tác và hiển thị, với các phương pháp thể hiện đồ họa các dữ liệu với số chiều, số lượng ít. Một số thí dụ về ứng dụng EDA:

¹ EDA: Exploratory Data Analysis: phân tích thăm dò dữ liệu.

- Dùng *thể hiện đồ họa* như khoanh bánh cho phép nhận biết tỉ lệ, có giá trị thực kèm theo; ứng dụng trong bệnh viện quân đội, Anh, 1958;
- Một *loạt các điểm*, cho phép thể hiện độ phì nhiêu của cây cối. Phương pháp phân tích thành phần chính với cách thể hiện này cho phép thấy kết quả đối sánh của một số đối tượng, 1987;
- Dùng *thể hiện không gian* trong ứng dụng mạng điện thoại, 1995.

6.1.4.2. Mô hình hóa mô tả

Đích của mô hình mô tả là mô tả tất cả các dữ liệu, hay mô tả quá trình sinh dữ liệu. Chẳng hạn những mô tả có mô hình là (i) phân phối xác suất của dữ liệu, tức đánh giá mật độ; (ii) phân đoạn không gian thành các nhóm, tức phân cụm và phân đoạn; (iii) mô hình mô tả quan hệ giữa các biến, tức mô hình hóa về sự phụ thuộc.

Việc phân đoạn là nhóm các bản ghi tương tự. Đích của phân đoạn là các nhóm đồng nhất. Số các nhóm do người nghiên cứu chọn, nên không có con số *đúng*. Ngược lại, trong phân tích cụm, đích là phát hiện nhóm *tự nhiên* của dữ liệu. Việc mô hình hóa mô tả đã được dùng theo vài cách:

- Phân đoạn trong tiếp thị sản phẩm, chia khách hàng thành các nhóm đồng nhất để cung cấp hàng, Wedel, Kamakura, 1998;
- Phân tích cụm trong nghiên cứu tâm lí trị liệu để phân lớp bệnh, Kendell, 1971;
- Các kĩ thuật phân cụm dùng trong phân tích dài hạn khi hậu tại bắc bán cầu, Cheng, Wallace, 1993.

6.1.4.3. Mô hình hóa dự đoán: phân lớp và hồi quy

Mục đích của việc này là xây dựng mô hình cho phép dự đoán giá trị của một biến theo các giá trị đã biết của biến khác.

Trong phân lớp, biến được dự đoán thuộc loại phạm trù, hạng mục; còn với hồi quy là biến định lượng.

Khái niệm *dự đoán* sử dụng ở đây không nhất thiết kèm yếu tố thời gian vô tận. Khái niệm *dự đoán* và *mô tả* trong công tác mô hình hóa phân biệt: (i) việc dự đoán nhằm vào biến duy nhất; (ii) việc mô tả xem xét tổng thể nhiều biến. Thí dụ về các mô hình dự đoán:

- Hệ thống SKICAT, Fayyad, 1996 dùng thể hiện cấu trúc cây trong việc học theo cây phân lớp, khi phân lớp các ngôi sao thiên hà;
- Công ty AT&T phát triển hệ thống dò đặc tính của trên 350 triệu điện thoại ở Mỹ, Cortes, Pregibon, 1998.

6.1.4.4. Phát hiện các khuôn mẫu và luật

Ba nhiệm vụ trên liên quan đến mô hình; nhiệm vụ này liên quan đến *khuôn mẫu*. Chẳng hạn theo dõi, đánh dấu các hành vi sai sót bằng cách phát hiện các vùng trong không gian có các giao tác khác biệt với vùng khác. Ngoài ra nhiệm vụ này phát hiện tổ hợp các thành phần xảy ra với tần suất cao trong các giao tác cơ sở dữ liệu. Vấn đề này được quan tâm nhiều trong khai phá dữ liệu, với kĩ thuật thuật toán hóa dựa trên luật kết hợp.

Điều có ý nghĩa ở đây là nhà thống kê quan tâm đến phát hiện bên lề, chỉ ra các hành vi bất thường trong ngữ cảnh của thay đổi bình thường. Điều này là phức tạp khi không gian có số chiều lớn. Người ta thấy tri thức lĩnh vực và diễn giải của con người không mấy tác dụng đối với vấn đề này. Một số thí dụ về phát hiện luật và khuôn mẫu:

- Huấn luyện chơi bóng cần các thông tin liên quan đến nhau; thể hiện qua hệ thống tìm luật, Bhandari, 1997;
- Áp dụng thuật toán học luật để phát hiện đặc tính hành vi vi phạm trong dịch vụ điện thoại, Fawcett, Provost, 1997.

6.1.4.5. Tìm kiếm nội dung

Khi người dùng có một khuôn mẫu, họ cần tìm các khuôn mẫu tương tự trong bộ dữ liệu. Đối với văn bản, khuôn mẫu là tập các từ khó; người dùng tìm các tệp văn bản. Đối với hình ảnh, người dùng có ảnh mẫu, khung xương ảnh, hay mô tả về ảnh; người ta cần tìm các ảnh tương tự trong kho các ảnh. Một số thí dụ:

- *Tìm kiếm trên Web* như Google, Brin, Page, 1998, đánh giá độ quan trọng của trang Web;
- Công ty IBM phát triển hệ thống QBIC¹, cho phép người dùng *tìm có tương tác* trong cơ sở dữ liệu ảnh theo câu hỏi mô tả nội dung, như màu, bề mặt, thông tin vị trí tương đối.

Các nhiệm vụ trên có thể không tách biệt rõ ràng. Khái niệm về *tương tự*, *khoảng cách* được dùng trong nhiều nhiệm vụ. Khái niệm hàm đích cũng được dùng, nhưng được hiểu theo cách khác nhau trong các nhiệm vụ. Các nhiệm vụ đều cần đến cấu trúc mô hình, khuôn mẫu, hay cấu trúc khác nếu cần thiết.

6.1.5. Các thành phần của thuật toán khai phá dữ liệu

Thuật toán được đề xuất để giải các nhiệm vụ đặt ra phần trước, được chia thành bốn thành phần cơ bản.

1. *Cấu trúc mô hình và khuôn mẫu*: xác định cấu trúc vốn có hay dạng chức năng từ các dữ liệu;
2. *Hàm đích*: kiểm định chất lượng của mô hình thích hợp;
3. *Phương pháp tìm kiếm và tối ưu*: tối ưu hàm đích và tìm trong mô hình và các cấu trúc khuôn mẫu khác;
4. *Chiến lược quản trị dữ liệu*: truy cập dữ liệu hiệu quả khi tối ưu/tìm kiếm.

Nội dung về cấu trúc mô hình, khuôn mẫu đã được trình bày trong phần trước. Tiếp theo là giải thích về ba thành phần còn lại.

¹ Query By Image Content: hỏi theo nội dung hình ảnh.

6.1.5.1. Các hàm đích

Định nghĩa: Hàm đích là hàm cho phép lượng hóa mức độ phù hợp của mô hình hay các tham số đối với bộ dữ liệu.

Thực tế cho thấy khó có thể xác định đúng, chính xác mức độ sử dụng của mô hình. Do vậy hàm đích đưa ra thường là tổng quát. Hàm đích có vai trò so sánh các mô hình, cho phép xác định tập tham số nào là tốt. Người ta thường dùng các hàm với mục đích kiểm tra mức độ giống nhau, sai số ít, tỉ lệ phân lớp cao.

Tuy về lí thuyết, hàm đích được xác định trước, một số tham số được chỉnh lí qua quá trình thực hiện, thử nghiệm. Việc điều chỉnh các tham số nhằm đảm bảo hàm đích phù hợp hơn.

6.1.5.2. Các phương pháp tìm kiếm và tối ưu

Hàm đích cho phép đo mức độ phù hợp của mô hình hay khuôn mẫu. Thông thường các mô hình hay khuôn mẫu được mô tả theo thuật ngữ của một cấu trúc, với các giá trị tham số chưa biết. Mục đích của việc tối ưu và tìm kiếm là xác định cấu trúc và giá trị tham số, để hàm đích đạt giá trị tối ưu.

Việc phát hiện các khuôn mẫu tốt trong số nhiều khuôn mẫu thường cần đến tìm kiếm tổ hợp, thường được giải bằng phương pháp may rủi. Phương châm ở đây là xác định khuôn mẫu chấp nhận được, tốt đối với cặp mô hình, khuôn mẫu và dữ liệu; chưa chắc tốt đối với trường hợp khác.

6.1.5.3. Các chiến lược quản trị dữ liệu

Người ta có cách tổ chức, lưu trữ, chỉ số hóa... và phương pháp truy cập dữ liệu. Nhiều thuật toán phân tích dữ liệu trong thống kê, học máy được phát triển với giả thiết dữ liệu sẽ được truy cập nhanh và hiệu quả trong bộ nhớ RAM, chưa đề cập bộ nhớ ngoài¹. Nhiều thuật toán phân tích dữ liệu chưa có đặc tả rõ

¹ RAM: Random Access Memory: bộ nhớ truy cập ngẫu nhiên; Bộ nhớ ngoài: băng đĩa, trống từ.

về chiến lược quản trị dữ liệu. Thuật toán mới chỉ thử nghiệm trên bộ dữ liệu nhỏ, rồi thực hiện trên bộ dữ liệu lớn trên bộ nhớ ngoài.

Lĩnh vực cơ sở dữ liệu đề cập các phương pháp chỉ số hóa, cấu trúc dữ liệu, thuật toán hỏi để lưu trữ và tìm kiếm dữ liệu hiệu quả.

6.1.6. Vai trò tương tác trong phân tích thống kê và khai phá dữ liệu

Các kĩ thuật phân tích thống kê chưa đủ cho khai phá dữ liệu, nhất là khi dữ liệu nhiều. Tuy nhiên thống kê là quan trọng trong bất kì quá trình khai phá dữ liệu nào.

Trong bộ dữ liệu lớn, không thể rút ngay ra tri thức, mà trước hết cần làm rõ dữ liệu. Một số kĩ thuật như xử lí thô trước khi đưa dữ liệu vào thuật toán, điển một số dữ liệu thiếu, chưa chính xác, tính toán trung gian... là cần thiết.

Phương pháp thu thập dữ liệu cũng ảnh hưởng đến khai phá dữ liệu, ở chỗ dữ liệu dễ bị sai lệch, do thiếu, đứt đoạn dữ liệu trong quá trình thu thập.

Mặt khác, không đòi hỏi mô hình, khuôn mẫu hoàn toàn phù hợp với bộ dữ liệu. Dữ liệu thu được mới chỉ là một phần của tất cả dữ liệu có thể có; bản thân chúng cũng chưa là đại diện đúng nhất cho thế giới cần được thể hiện. Vậy nên, người ta không tìm các mô hình, khuôn mẫu khớp với bộ dữ liệu, mà đưa ra mô hình, khuôn mẫu có mức tổng quát cao đối với nhiều bộ dữ liệu.

Khác nhau giữa các ứng dụng thống kê cổ điển với khai phá dữ liệu là lượng dữ liệu. Đối với thống kê, khái niệm *lớn* chỉ là vài trăm, thì đối với khai phá dữ liệu, *lớn* ứng với vài triệu.

Vấn đề số lượng lớn của dữ liệu có thể được khắc phục nhờ các phương pháp thích nghi, tổng hợp dữ liệu, hay phương pháp lấy mẫu. Việc lấy mẫu chỉ thích hợp khi tìm mô hình, chứ không thích hợp để tìm khuôn mẫu.

Người ta còn khắc phục mức độ lớn của các biến. Theo quan điểm của phân tích thống kê bình thường, dữ liệu gồm các tệp dữ

liệu phẳng, dễ hình dung. Thực tế, dữ liệu còn được lưu trữ trên nhiều máy, có nhiều chiều; điều này đòi hỏi khai phá dữ liệu chú trọng nhiều về (i) tổ chức dữ liệu; (ii) thuật toán trên dữ liệu.

Một vấn đề khó khăn như độ lớn của dữ liệu tạo ra là thiếu một số thuộc tính dữ liệu. Khai phá dữ liệu thường là giai đoạn sau phân tích thống kê. Giai đoạn đầu, các dữ liệu được thu thập theo mục đích đáp ứng yêu cầu của phân tích thống kê. Giai đoạn sau thiếu các thông tin mà phân tích thống kê không yêu cầu. Lúc này, người ta cần dữ liệu bình thường, thể hiện các tình huống, chứ không phải các dữ liệu ngẫu nhiên như phân tích thống kê đã lấy.

6.1.7. Thuật ngữ trong quá trình khai phá dữ liệu

Người ta đã từng sử dụng thuật ngữ *đánh bắt, dò tìm, nạo vét*¹ trước khi dùng thuật ngữ khai phá dữ liệu. Từ năm 1960, người ta đã có nhu cầu về mô hình hóa khối lượng lớn các dữ liệu. Hai yếu tố được người ta quan tâm là (i) độ phức tạp của mô hình; và (ii) kích thước của tập các mô hình chấp nhận được.

Nếu các mô hình là mềm dẻo, tức phù hợp tương đối với bộ dữ liệu, người ta dễ chọn được mô hình khớp với dữ liệu. Điều này khả thi do các mô hình là tổng quát. Mặt khác, nếu người dùng cần mô hình thật khớp với dữ liệu, có thể tổng hợp một cách chính xác trên các dữ liệu, để được bộ dữ liệu phù hợp. Thực ra, việc cố đạt được mô hình thật khớp về lý thuyết, chưa thật có giá trị đối với ứng dụng.

Về thuật ngữ *khám phá*², các nhà thống kê học đã thấy không thuận, khi nó có xu hướng mờ rộng hơi quá trong trường hợp bộ dữ liệu lớn, không gian các cấu trúc xem xét đủ lớn. Vấn đề xảy ra đối với tìm khuôn mẫu, chứ đối với mô hình thì không; khám phá khuôn mẫu cần số lượng tương đối ít các trường hợp.

¹ Fishing: đánh bắt; Snooping: dò tìm; Dredging: nạo vét (dữ liệu).

² Discovery: khám phá.

Để giải quyết vấn đề trên, không thể chỉ với kĩ thuật như tách bộ dữ liệu để phát hiện khuôn mẫu trên tập con... mà cần xét đến ngữ nghĩa của dữ liệu.

Vấn đề đặt ra trong phần trên là trích các thông tin đáng giá trong bộ dữ liệu có số lượng lớn. Khai phá dữ liệu là lĩnh vực mới, cho phép *tìm kiếm* trong cơ sở dữ liệu, tổng hợp dữ liệu, phát hiện khuôn mẫu.

Các dữ liệu có thể được *phân tích thống kê* theo nhiều cách. Nhiều cấu trúc, khuôn mẫu được khai phá dữ liệu chỉ ra. Quá trình này đòi hỏi các thuật toán trong thống kê và công nghệ thông tin.

6.2. Về các thuật toán khai phá dữ liệu theo quan điểm hệ thống

6.2.1. Giới thiệu về thuật toán

Sau đây là định nghĩa về *thuật toán khai phá dữ liệu*, để biết nó nhằm đến nghĩa nào, có thành phần gì.

Định nghĩa: Thuật toán khai phá dữ liệu là thủ tục xác định tốt nhập dữ liệu và xuất ra dạng mô hình hay dạng mẫu.

Xác định tốt nhằm ý thủ tục có thể chuyển sang mã chính xác bằng tập hữu hạn các quy tắc. Để là thuật toán, thủ tục cần kết thúc sau số hữu hạn bước, và cho ra kết quả. Ngược lại, *phương pháp tính*¹ có tất cả các tính chất như thuật toán, ngoài việc không đảm bảo kết thúc. Thuật toán đòi hỏi nhiều thực hiện cụ thể; trong khi phương pháp tính được mô tả trừu tượng hơn.

Đặc tả thuật toán khai phá dữ liệu để giải nhiệm vụ cụ thể yêu cầu xác định các thành phần thuật toán:

1. *Nhiệm vụ khai phá dữ liệu*, như hiển thị, phân lớp, phân cụm, hồi quy... về bản chất, các loại thuật toán khác nhau dùng cho nghiên cứu khác nhau;

¹ Computational method: phương pháp tính.

2. *Cấu trúc, dưới dạng hàm, của mô hình hay mẫu khít với dữ liệu*, như mô hình hồi quy tuyến tính, mô hình phân cụm phân cấp... Cấu trúc xác định biên người ta xấp xỉ hay học. Trong biên này, dữ liệu hướng dẫn người ta đến mô hình hay mẫu cụ thể;
3. *Hàm đích dùng để kiểm chứng chất lượng của các mô hình hay các mẫu khít với dữ liệu quan sát*, như sai số phân lớp, sai số bình phương... Quan trọng là hàm đích phản ánh tiện ích thực tế tương đối của việc lấy tham số khác nhau với mô hình hay mẫu. Ngoài ra, hàm mẫu là căng đối với việc học và tổng quát hóa. Nó có thể dựa trên chỉ việc khớp với phần chính, chẳng hạn như cách dữ liệu quan sát mô tả rõ mô hình;
4. *Phương pháp tìm tối ưu dùng để tìm tham số và cấu trúc*, như các thủ tục tính toán và các thuật toán dùng để tìm hàm đích cực đại đối với mô hình hay mẫu cụ thể. Phần này sẽ có các phương pháp tính để tối ưu hàm đích và tìm ra tham số. Nếu cấu trúc mô hình là cấu trúc cố định đơn, như hàm đa thức bậc k của đầu vào, việc tìm kiếm được đưa về không gian tham số để tối ưu hàm đích theo dạng cấu trúc cố định này. Nếu cấu trúc mô hình, hay mẫu, gồm tập, hay họ, các cấu trúc khác nhau, sẽ tìm trên cả các cấu trúc và không gian tham số liên quan. Việc tối ưu hóa và tìm kiếm là vấn đề chính của thuật toán khai phá dữ liệu;
5. *Kỹ thuật quản lý dữ liệu để lưu trữ, chỉ số hóa và tìm dữ liệu*. Nhiều thuật toán học máy và thuật toán thống kê không xác định kỹ thuật quản lý dữ liệu, mà giả thiết rằng bộ dữ liệu đủ nhỏ chứa được trên bộ nhớ chính, và giá truy cập thấp so với giá tính toán. Sự thực thì dữ liệu đa số trên bộ nhớ phụ. Vây cần có thuật toán phù hợp.

Bảng sau mô tả ba thuật toán khai phá dữ liệu nổi tiếng chia theo các thành phần thuật toán.

Bảng 6.1. Ba thuật toán nổi bật của khai phá dữ liệu

	CART	Truyền ngược	Apriori
Nhiệm vụ	Phân lớp và hồi quy	Hồi quy	Phát hiện mẫu luật
Cấu trúc	Cây quyết định	Mạng thần kinh (các hàm phi tuyến)	Luật kết hợp
Hàm đích	Hàm mất mát được kiểm chứng chéo	Sai số bậc hai	Trợ giúp / đúng
Phương pháp tìm kiếm	Tìm tham lam trên cấu trúc	Hạ dốc theo tham số	Tìm rộng có cắt
Kĩ thuật quản lí dữ liệu	Chưa xác định	Chưa xác định	Quét tuyến tính

Việc xác định cấu trúc của mô hình, hay mẫu, và hàm đích thường không đồng thời với quá trình xác lập bài toán khai phá dữ liệu theo cách lấy người dùng làm trung tâm. Sau khi có dữ liệu, cấu trúc mô hình, hay mẫu, và hàm đích, vấn đề tiếp theo là tối ưu hàm đích. Việc này cần nhiều tính toán. Thực tế cho thấy cần lặp lại quá trình. Do vậy cốt yếu của việc khai phá dữ liệu nằm ở các phương pháp tính toán dùng để thể hiện các thành phần quản trị dữ liệu và tìm kiếm.

Mô tả dựa trên thành phần vừa nêu đảm bảo khung mức tổng quát mức cao, cả về phân tích lẫn tổng hợp của các thuật toán khai phá dữ liệu.

- *Theo quan điểm phân tích*, việc mô tả thuật toán theo thuật ngữ thành phần sẽ làm sáng tỏ vai trò của thành phần, để so sánh với thuật toán khác;
- *Theo quan điểm tổng hợp*, bằng cách kết hợp các thành phần theo cách đa dạng, có thể dựng được các thuật toán khai phá dữ liệu có các tính chất khác nhau.

Chủ đề chính ở đây là quan điểm theo thành phần về các thuật toán khai phá dữ liệu đảm bảo ngôn ngữ chi tiết và cấu trúc để mô tả, phân tích và tổng hợp các thuật toán khai phá dữ liệu.

Người ta hạn chế ở trường hợp có dạng đơn của mô hình, hay mẫu, như cây, đa thức..., hơn là trường hợp có dạng kép của các cấu trúc mô hình đối với cùng bài toán. Quan điểm thành phần có thể được tổng quát hóa để quản lí các trường hợp, nhưng điển hình là quản lí các hàm đích, phương pháp tìm kiếm và kĩ thuật quản trị dữ liệu.

6.2.2. *Thí dụ về thuật toán CART xây dựng bộ phân lớp theo cây*

Để thấy rõ ý tưởng xem xét thuật toán theo các thành phần, trước tiên xét thuật toán CART¹, là thủ tục thống kê sử dụng rộng rãi để phân lớp và hồi quy trong với cấu trúc dựa trên cây. Để đơn giản, chỉ xét khía cạnh phân lớp; đó là khớp vectơ nhập x với đầu ra y về hạng mục, hay lớp.

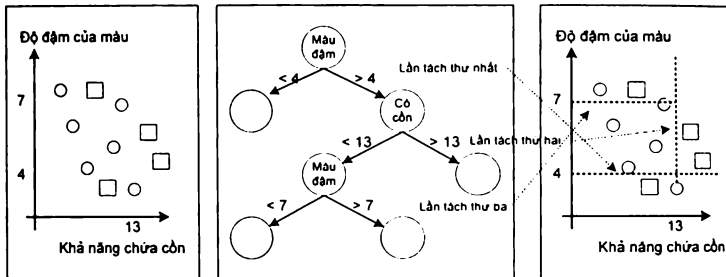
Theo nghĩa các thành phần, thuật toán CART được nhìn nhận như các bước *thuật toán* (i) nhiệm vụ; ở đây là dự đoán, phân lớp; (ii) cấu trúc mô hình; ở đây là cấu trúc cây; (iii) hàm đích; ở đây là hàm tổn thất kiểm chứng chéo; (iv) phương pháp tìm kiếm; ở đây sử dụng thuật toán tìm tham lam; (v) phương pháp quản trị dữ liệu; hiện chưa xác định.

Khía cạnh phân biệt cơ bản của thuật toán CART là cấu trúc mô hình được dùng; tức cây phân lớp. Mô hình cây phân lớp gồm phân cấp của các quyết định nhị phân một biến. Mỗi nút trong cây xác định kiểm tra đúng/sai trên một biến, dùng ngưỡng có giá trị biến thực hay nguyên và tập con của giá trị biến hạng mục.

Cấu trúc cây có được do dữ liệu, chứ chưa được xác định trước. Việc xác định trước lại là vấn đề mà khai phá dữ liệu quan tâm. CART hoạt động theo cách chọn ra biến tốt nhất để tách dữ liệu thành hai nhóm từ nút gốc. Có thể dùng bất kì tiêu chí tách, cho phân đoạn hiệu quả các dữ liệu từ một nút trong cây thành hai

¹ CART: Classification And Regression Trees: phân lớp và cây hồi quy.

cây con, có nhãn và đồng nhất. Kích thước cây sau cùng là kết quả của quá trình *cắt, xén*¹.



Hình 6.1. Phân tách các miền quyết định

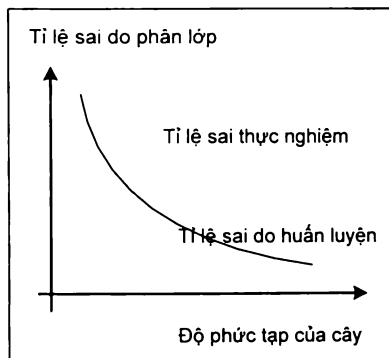
Dạng phân cấp của cấu trúc cây phân biệt thuật toán như CART với các thuật toán phân lớp dựa trên các cấu trúc không dùng cây, tức mô hình dùng tổ hợp tuyến tính các tất cả các biến để xác định biên quyết định trong không gian dữ liệu nhập. Cấu trúc cây sử dụng cho phân lớp có thể được xử lý với dữ liệu trộn, tức cả phạm trù, hạng mục và dữ liệu thực, do mỗi nút trong cây phụ thuộc vào kiểm tra một biến đơn. Ngoài ra, do CART xây dựng cây dùng biến đơn tại một thời điểm, có thể sử dụng nhiều biến cho bài toán. Nói cách khác, năng lực thể hiện của cấu trúc cây hơi thô: miền quyết định để phân lớp bị hạn chế ở hình chữ nhật, đường biên song song với trục dữ liệu.

Hàm đích dùng để xác định chất lượng của cấu trúc cây khác nhau, thường được dùng là hàm tổn thất do phân lớp sai, được xác định là $\sum_{i=1}^n C(y(i), \hat{y}(i))$, trong đó $C(y(i), \hat{y}(i))$ là mất mát, là dương, phải chịu khi nhận phân lớp đối với vectơ dữ liệu thứ i , là $y(i)$, do cây nhận là $\hat{y}(i)$. Nhìn chung C được xác định như ma trận $m \times m$, m là số lớp. Để đơn giản, người ta giả thiết lỗi là 1 nếu $\hat{y}(i) \neq y(i)$, ngược lại là 0. Đây là hàm lỗi 0 - 1, hay tỉ lệ phân lớp sai nếu người ta chuẩn hóa tổng trên bằng cách chia cho n .

¹ Pruning: cắt, xén.

CART dùng kĩ thuật gọi là kiểm chứng chéo, để ước lượng hàm tổn thất do phân lớp sai. Về cơ bản phương pháp này chia dữ liệu thành tập con để xây dựng cây, rồi ước lượng tỉ lệ phân lớp sai trên tập kiểm chứng còn lại. Việc chia, hay phân đoạn này được lặp nhiều lần trên tập con khác nhau, và tỉ lệ phân lớp sai được lấy trung bình để được ước lượng kiểm chứng chéo về cây với kích thước cụ thể sẽ thực hiện đối với cây mới, chưa biết trước tốt ra sao. Kích thước của cây có ước lượng tổn thất thấp nhất được chọn như kích thước phù hợp đối với mô hình cây cuối cùng.

Kiểm chứng chéo cho phép CART ước lượng hiệu năng của mô hình cây bất kì trên dữ liệu chưa dùng để kiến thiết nên cây; tức là nó đảm bảo ước lượng hiệu năng tổng quát hóa. Điều này là đòi hỏi nhiều công sức trong thủ tục phát triển cây, do tỉ lệ phân lớp sai trên dữ liệu huấn luyện, tức dữ liệu tạo nên cây, thường bị giảm đi để đơn giản cây, vậy nên sai số dữ liệu huấn luyện không đủ cho thấy cách cây làm việc với dữ liệu mới.



Hình 6.2. Đồ thị sai số

Hình trên cho thấy tỉ lệ sai do huấn luyện giảm đơn điệu. Tỉ lệ sai do thử đầu tiên giảm. Cây nhỏ không đủ năng lực dự đoán mức chính xác. Cây phức tạp hơn sẽ gây lỗi thử tăng, ứng với việc thuật toán không còn phù hợp với dữ liệu; việc thêm nút chỉ gây sai lệch ngẫu nhiên trong dữ liệu huấn luyện, và không hợp lí với nhiệm vụ dự đoán. Đích của thuật toán CART là tìm ra cây sát với

kích thước cây tối ưu; cố gắng tìm ra mô hình đủ phức tạp để nắm bắt cấu trúc đã có, nhưng không để quá. Đối với bộ dữ liệu quá lớn, người ta có thể cố tách dữ liệu một cách đơn giản thành bộ dữ liệu huấn luyện và bộ dữ liệu kiểm chứng và giám sát hiệu năng trên bộ dữ liệu kiểm chứng.

Việc sử dụng hàm đích kiểm chứng chéo phân biệt CART với hầu hết các thuật toán khai phá dữ liệu khác dựa trên mô hình cây. Thuật toán C4.5, thuật toán như CART đã dùng rộng để xây dựng cây phân lớp, kiểm chứng các cấu trúc cây riêng bằng cách thêm vào theo kinh nghiệm tỉ lệ sai số ước lượng trên bộ dữ liệu huấn luyện để tính xấp xỉ tỉ lệ sai do thừa, nhằm chỉnh lí việc tỉ lệ sai do huấn luyện nói chung thấp hơn *tỉ lệ sai ra ngoài mẫu*¹. Tỉ lệ sai được chỉnh sẽ được dùng trong giai đoạn cắt, xén cây, để tối ưu hàm đích.

CART dùng phương pháp tìm kiếm tham lam để xác định các cấu trúc cây ứng cử viên bằng cách nối cây, đệ quy từ nút gốc, rồi xén dần các nhánh đặc biệt của cây lớn. Phương pháp tìm may rủi được điều khiển bởi không gian tìm kiếm lớn một cách tổ hợp, tức không gian các tất cả các cấu trúc cây nhị phân, và thiếu phương pháp dò vết để phát hiện cây tối ưu đơn liên quan đến hàm đích. Kinh nghiệm cho thấy phương pháp tìm tham lam tốt như phương pháp lí tưởng; do vậy nó được thực hiện trong các thuật toán thực hành tìm trên cây.

Về quản trị dữ liệu, CART ngầm giả thiết rằng dữ liệu được lưu tại bộ nhớ trong. Để tiện cho CART, rất ít thuật toán nêu cách tổ chức dữ liệu khi bộ dữ liệu lớn. Đối với một số thuật toán, dùng hướng dẫn về quản trị dữ liệu là thuận lợi, theo khối chương trình.

Đối với thuật toán cây, mô hình, hàm đích và phương pháp tìm kiếm là phức tạp tới mức cần xem xét cách tổ chức dữ liệu. Cấu trúc cây được tách đoạn đệ quy thành các tập con. Với cách tổ chức chưa tốt, cần lặp đi lặp lại động tác với bộ nhớ ngoài.

¹ Out - of - sample error rate: tỉ lệ sai ngoài mẫu.

Để nhìn nhận ngắn gọn về CART, lưu ý rằng thuật toán gồm (i) cấu trúc mô hình cây; (ii) hàm đích kiểm chứng chéo; (iii) tìm tham lam trên cấu trúc cây theo hai pha, phát triển và xén. Theo cách này, CART là dễ hiểu khi biết được ý chính. Người ta cần phát triển các thuật toán thay thế cho cùng một cấu trúc cây, hàm đích kiểm chứng chéo, và kĩ thuật tìm ứng với CART, nhưng chuyên biệt cho ứng dụng theo cài đặt cụ thể. Đối với ứng dụng khai phá dữ liệu, chính thuật toán phù hợp có giá trị cao. Túc quyền lực của thuật toán như CART là ở các khái niệm cơ bản, chứ không phải yếu tố cài đặt.

6.2.3. Nhìn nhận sơ bộ về các thuật toán khai phá dữ liệu

Chủ đề quan tâm là nhiệm vụ khám phá bộ dữ liệu đặc biệt trong bộ dữ liệu; thuật toán khai phá dữ liệu được coi như các bước {*cấu trúc mô hình, hàm đích, phương pháp tìm kiếm, kĩ thuật quản trị dữ liệu*}. Tuy đơn giản vậy, nhưng nó có ứng dụng thực tế. Thứ nhất là có một loạt các thuật toán, do kết hợp các mô hình khác nhau các hàm đích khác nhau...

Độ phức tạp của không gian thuật toán được quản lí, bởi lẽ tuy có nhiều thuật toán, nhưng chỉ có số ít các giá trị cơ bản đối với mỗi thành phần thuật toán. Đặc biệt có nhiều mô hình, mẫu, xác định tốt cho bài toán hồi quy, phân lớp, hay phân cụm. Cũng có lớp tìm kiếm, phương pháp tối ưu tổng quát cho phép sử dụng. Thuật toán có thể dùng các khía cạnh xác định tốt của các thành phần, tức gắn với không gian thuật toán.

Cần có nhìn nhận ngắn, theo các thành phần thuật toán, đối với các thuật toán khai phá dữ liệu có ý nghĩa thực tế. Nó làm rõ thao tác trong thuật toán cụ thể bằng cách đưa về các thành phần cốt yếu. Mặt khác điều này cho phép so sánh các thuật toán với nhau, chẳng hạn CART với C4.5.

Dù nhiều vấn đề cần nhìn nhận, ở đây chỉ nhấn mạnh vào tính chất cơ bản của một thuật toán, tránh đề cập đến nhiều thuật

toán. Khi đề cập thuật toán khai phá dữ liệu, người ta thường nghĩ về các thành phần khớp với đặc tính của bài toán, hơn là thuật toán nền tổng quát. Có thể dùng phần mềm để tổ hợp các thành phần từ thư viện. Tuy nhiên đó mới chỉ là ý tưởng, chưa thực tế; người ta thường liệt kê danh sách các thuật toán, thay vì liệt kê các thành phần giúp cho việc tổng hợp. Tiếp cận *cẩm nang nấu ăn*¹ đôi khi nguy hiểm do người ít kinh nghiệm về khai phá dữ liệu không hiểu hết các hạn chế, và các giả thiết sau đó của thuật toán hợp đen sử dụng. Ngược lại, mô tả theo thành phần làm sáng tỏ bên trong hộp đen.

Phần tiếp theo sẽ xét sơ lược ba thuật toán theo thuật ngữ mô tả các bước, các thành phần.

6.2.3.1. Máy hồi quy và phân lớp nhiều tầng

Máy nhiều tầng MLP² là mô hình được sử dụng nhiều nhất để phân lớp tổng quát theo mô hình thần kinh nhân tạo. Cấu trúc MLP đảm bảo khớp phi tuyến các vectơ x giá trị đầu vào thực với vectơ đầu ra y . MLP dùng như mô hình phi tuyến cho bài toán hồi quy, hay phân lớp qua việc diễn giải các đầu ra. Ý tưởng cơ bản là:

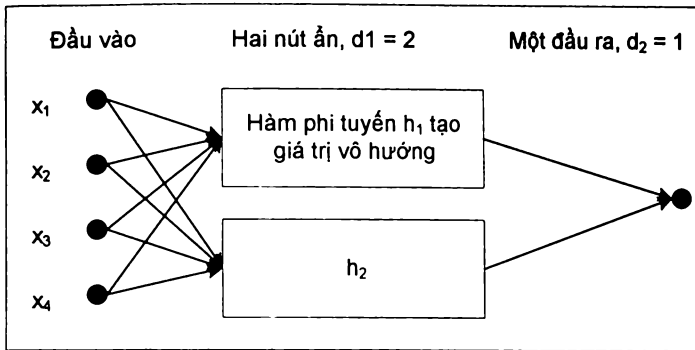
1. Vectơ giá trị vào p được nhân với ma trận trọng số $p \cdot d_1$, và giá trị d_1 tạo nên được chuyển riêng từng cái bằng hàm phi tuyến để thu được d_1 đầu ra, *nút ẩn*;
2. Các giá trị d_1 nhân với ma trận trọng số $d_1 \cdot d_2$, tức trọng số tầng khác, và các giá trị d_2 thu được nhờ hàm phi tuyến;
3. Các giá trị d_2 được dùng như giá trị đầu ra của mô hình, hoặc chuyển qua tầng khác và chuyển hóa phi tuyến theo mô hình.

Thuật ngữ *máy*³ có từ mô hình ban đầu, đưa ra năm 1960, có tầng đơn và phi tuyến theo ngưỡng.

¹ Cookbook: cẩm nang nấu ăn.

² Multiplayer perceptron

³ Perceptron



Hình 6.3. Sơ đồ mạng nhận biết nhiều tầng, tức mạng thần kinh, với hai nút ẩn ($d_1 = 2$) và một nút ra ($d_2 = 1$)

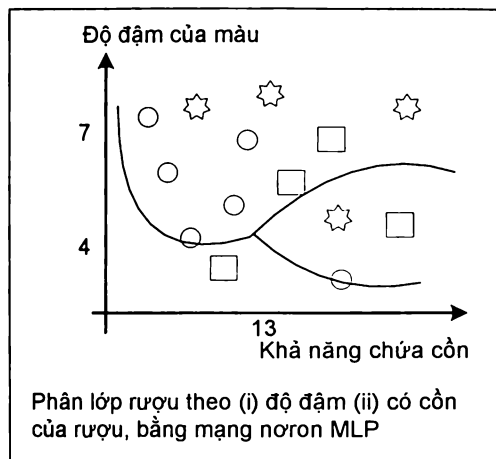
MLP có một *tầng ẩn*, sinh ra $s_1 = \sum_{i=1}^4 \alpha_i x_i, s_2 = \sum_{i=1}^4 \beta_i x_i$ từ tầng thứ nhất, trọng số α, β . Các s được chuyển hóa nhờ hàm phi tuyến h_1, h_2 . Hàm phi tuyến thường được dùng là $h_i = h(s_i) = \frac{1}{1 + e^{-s_i}}$.

Các giá trị h được gán trọng số để tính $y = \sum_{i=1}^2 w_i h_i$. Có thể dùng hàm phi tuyến đối với y . Vậy y là hàm phi tuyến của vector đầu vào x . Các tham số mô hình là α_i, β_i và hai trọng số w_i . Nói chung khi có p đầu vào, với MLP tầng đơn có n_h nút ẩn, một đầu ra đơn, có $(p + 1) \cdot n_h$ tham số cần được ước lượng qua dữ liệu. Hệ thống sẽ chậm nếu dùng nhiều tầng ẩn.

Biến y là vô hướng, tức $d_2 = 1$, và có thể đưa về giá trị giữa 0, 1, nên người ta dùng MLP để phân lớp, với ngưỡng 0.5 nếu thuộc tập thứ nhất, thứ hai. Do bản chất phi tuyến của mô hình, ngưỡng khác nhau để nhận biết các lớp có thể phi tuyến. Mức phi tuyến có thể cao, vậy không có dạng đơn giản để mô tả mô hình mạng thần kinh.

Nhìn nhận thu gọn thuật toán học MLP theo các bước thuật toán (i) *nhiệm vụ*: dự đoán, là phân lớp hay hồi quy; (ii) *cấu trúc*: các tầng chuyển hóa phi tuyến các đầu vào với tổng có trọng số; (iii) *hàm đích*: giảm từng bước qua việc chọn ngẫu nhiên các giá

trị tham số khởi đầu; (iv) *kỹ thuật quản trị dữ liệu*: trực tuyến hay theo lô.



Hình 6.4. Biên quyết định mà mô hình mạng thần kinh thiết lập

Nét nổi bật của thuật toán là nhiều tầng, bản chất phi tuyến của cấu trúc mô hình. Giá trị (i) đầu ra y là hàm phi tuyến của các đầu vào x ; (ii) các tham số, tức trọng số, trong hàm đích đều phi tuyến. Mạng thần kinh khác với các dạng hàm đa thức tuyến tính truyền thống dùng trong hồi quy và mô hình phân lớp theo cây.

Tổng sai số bình phương SSE^1 là hàm đích thông dụng cho MLP được xác định qua $S_{SSE} = \sum_{i=1}^n (y(i) - \hat{y}(i))^2$, trong đó $y(i)$, $\hat{y}(i)$ là giá trị cuối thực và đầu ra của mạng thần kinh cho điểm dữ liệu thứ i ; và $\hat{y}(i)$ là hàm của vector nhập $x(i)$ và θ là tham số của MLP. Đôi khi giả thiết sai số bình phương là hàm đích; hàm này có thể dùng cho mô hình mạng thần kinh. Thực tế có thể dùng hàm của tham số mô hình, nên bất kì hàm đích nào cũng được dùng như cơ sở của phương pháp tìm giảm nhanh như truyền ngược.

¹ Sum of squared error: tổng sai số bình phương.

Huấn luyện mạng thần kinh gồm việc tối thiểu S_{SSE} ; coi nó như hàm của tham số θ chưa biết. Cho mỗi $y(i)$ là hàm phi tuyến của tham số θ , hàm đích S_{SSE} cũng là hàm phi tuyến cao của θ . Vậy nên không có lời giải theo dạng đóng khi tìm tham số θ để cực tiểu S_{SSE} . Ngoài ra, do có nhiều cực trị địa phương trên mặt S_{SSE} , nên việc huấn luyện mạng thần kinh trở nên bài toán tối ưu đa biến không tầm thường. Kỹ thuật tìm cực bộ lặp cần thiết để tìm điểm cực tiểu địa phương vừa ý.

Phương pháp huấn luyện ban đầu dùng cho MLP, tức lan truyền ngược, là phương pháp tối ưu tương đối đơn giản. Nó thực hiện giảm nhanh hàm đích, tức tổng sai số bình phương, trong không gian tham số, khi giải bài toán tối ưu phi tuyến bằng cách giảm giá trị cực tiểu địa phương với điểm khởi đầu ngẫu nhiên trong không gian tham số. Trong ngữ cảnh tổng quát, tồn tại họ phương pháp tối ưu giải các bài toán phi tuyến. Thường giả thiết giảm nhanh là phương pháp tối ưu duy nhất dùng để huấn luyện mạng MLP, nhưng trong thực tế có các kỹ thuật tối ưu phi tuyến mạnh như *kỹ thuật độ dốc liên hợp*¹.

Về quản trị dữ liệu, mạng thần kinh có thể được huấn luyện hoặc trực tuyến, hoặc theo lô. Trực tuyến theo nghĩa cập nhật trọng số khi qua một điểm dữ liệu; theo lô theo nghĩa cập nhật trọng số sau khi xét hết tất cả các dữ liệu.

Một phân biệt quan trọng giữa MLP và cây phân lớp là thuật toán cây, như CART, tìm qua mô hình với độ phức tạp khác nhau theo kỹ thuật otomat. Ngược lại, không có thủ tục xác định cho MLP, tức chưa xác định được số tầng, số nút ẩn trong mô hình. Nhiều thuật toán cho phép thiết lập cấu trúc mạng một cách tự động, cả thuật toán phát triển dần các nút và trọng số, và cắt xén. Việc tăng nhanh cấu trúc mạng coi như bài toán cực trị địa phương; mạng tốt nhất với k nút ẩn có thể hoàn toàn khác trong không gian tham số với mạng tốt nhất với $k - 1$ tham số. Mặt khác,

¹ Conjugate gradient technique: kỹ thuật độ dốc liên hợp.

huấn luyện mạng lớn có thể tốn kém, đặc biệt khi cấu trúc mô hình lớn, tức có nhiều đầu vào p . Trong thực tế, các cấu trúc mạng thường được xác định bằng thủ tục *thử sai*¹ khi thêm thủ công các nút ẩn cho đến khi đạt hiệu quả mong muốn trên bộ dữ liệu kiểm chứng, tức bộ dữ liệu không dùng để huấn luyện.

Cách nhìn theo thành phần của MLP minh họa rằng tiếp cận tổng quan không khác ước lượng thống kê truyền thống và các kỹ thuật tối ưu. Nhiều kỹ thuật, như (i) kỹ thuật tích hợp thông tin về sự kiện tiên nghiệm Bayes vào hàm đích để lái trọng số về 0, để chính quy hóa mô hình; hay (ii) kỹ thuật dùng các thủ tục tối ưu nhiều biến lý tưởng hơn như kỹ thuật độ dốc liên hợp khi tìm trọng số, có thể được dùng trong các mô hình huấn luyện mạng. Vào những năm 80, khi các mô hình mạng thần kinh xuất hiện, liên hệ với khía cạnh thống kê chưa rõ ràng. Nhưng không nghi ngờ về việc tiếp cận mô hình thần kinh là ở bản chất nhiều tầng phi tuyến của cấu trúc mô hình.

6.2.3.2. Thuật toán Apriori đối với việc học luật kết hợp

Luật kết hợp là một trong các thể hiện thông thường nhất cho mẫu cục bộ trong khai phá dữ liệu. Phần này trình bày vài ý chính liên quan đến thuật toán Apriori, thuật toán sớm nhất trong việc phát hiện luật kết hợp. Một luật kết hợp là khẳng định xác suất về việc xảy ra đồng thời của vài sự kiện trong cơ sở dữ liệu, và được dùng cụ thể để làm chi tiết, cụ thể hơn về bộ dữ liệu giao tác.

Để đơn giản, giả sử tất cả các biến là nhị phân. Một luật kết hợp có dạng $IFA = 1 \text{ AND } B = 1 \text{ THEN } C = 1$ với xác suất p ; trong đó A, B, C là biến nhị phân và $p = p(C = 1 \mid A = 1, B = 1)$. Xác suất điều kiện p đôi khi được dùng như *chuẩn xác*, *độ tin cậy* của luật, và $p(A = 1, B = 1, C = 1)$ là *trợ giúp*². Cấu trúc mẫu hay cấu trúc luật dễ hiểu và đơn giản. Đích đến hình là tìm tất cả các luật đáp ứng

¹ Trial - and - error: thử sai.

² Accuracy: chuẩn xác; Confidence: tin cậy; Support: trợ giúp.

ràng buộc, độ chính xác lớn hơn p_a , trợ giúp hơn p_s . Các luật như vậy có cả dạng yếu của tri thức; chúng chính là rút gọn của các mẫu đồng thời trong dữ liệu quan sát, chứ không hẳn là khẳng định mạnh đặc trưng cho toàn bộ dân cư. Vậy, nghĩa của *luật* là ở diễn giải nhân quả; thuật ngữ *luật kết hợp* thực ra là dùng sai vì các mẫu này cùng quan hệ thừa kế, chứ không cần nhân quả.

Ý chung của phát hiện luật kết hợp xuất phát từ dữ liệu mua hàng. Các dữ liệu này thường được ghi trong cơ sở dữ liệu; mỗi quan sát gồm các mục mặt hàng, và các biến cho thấy mặt hàng được mua hay không. Có thể hình dung như dữ liệu trong ma trận dữ liệu n dòng ứng với n rõ hàng, p cột ứng với p mặt hàng. Ma trận có thể lớn, n là hàng triệu và p vài nghìn, và thưa do có rất ít hàng. Luật kết hợp là cách đơn giản phát hiện mẫu đơn giản trong dữ liệu theo tính toán hiệu quả.

Theo cách nhìn thu gọn, thuật toán khai phá dữ liệu điển hình đối với luật kết hợp có các thành phần (i) *nhiệm vụ*: mô tả, gồm việc kết hợp giữa các biến; (ii) *cấu trúc*: luật kết hợp, hay các mẫu, xác suất; (iii) *hàm đích*: các ngưỡng về chính xác và trợ giúp; (iv) *phương pháp tìm*: tìm có hệ thống, tìm rộng có cắt xén; (v) *kỹ thuật quản trị dữ liệu*: quét tuyến tính nhiều lần.

Hàm đích dùng trong tìm luật kết hợp đơn giản là hàm nhị phân. Có hai ngưỡng (i) p_s là biên thấp trợ giúp luật, tức $p_s = 0.1$ khi người ta muốn luật này bao phủ 10% dữ liệu; và (ii) p_a là biên thấp về chính xác, tức $p_a = 0.9$ khi người ta muốn ít nhất 90% là chính xác. Một mẫu lấy đích là 1 nếu thỏa mãn cả hai điều kiện ngưỡng; ngược lại 0. Đích là phát hiện ra tất cả các luật, mẫu, với đích 1.

Bài toán tìm kiếm là phức tạp khi các luật kết hợp có thể là theo lũy thừa, $O(p^{2p-1})$, đối với các biến nhị phân nếu chỉ quan tâm đến luật với tỉ lệ dương, tức $A = 1$, cả bên trái và phải. Nếu lưu ý đến bản chất của hàm đích, người ta có thể giảm thời gian trung bình của thuật toán. Nếu $p(A = 1) \leq p_s$ hay $p(B = 1) \leq p_s$ thì $p(A = 1,$

$B = 1) \leq p_s$. Dùng quan sát này khi tìm luật kết hợp theo cách (i) tìm các sự kiện riêng, như $A = 1$, có xác suất lớn hơn p_s ; (ii) sự kiện được gọi là *thường xuyên* nếu xác suất sự kiện lớn hơn ngưỡng trợ giúp p_s . Tất cả các cặp sự kiện thường xuyên là tập ứng cử thường xuyên với kích thước 2.

Trong trường hợp tổng quát với tập thường xuyên kích thước $k - 1$ sang tập kích thước k , người ta có thể cắt tập kích thước k bất kì có tập con $k - 1$ để nó không là tập thường xuyên với kích thước $k - 1$. Việc cắt xén này không tìm dữ liệu một cách trực tiếp, nhưng làm giảm lượng tính toán đáng kể.

Cho danh sách cắt xén của các tập ứng cử thường xuyên kích thước k , thuật toán quét tuyến tính cơ sở dữ liệu để xác định tập nào thực sự là thường xuyên. Tập thường xuyên được khẳng định kích thước k , nếu có, được kết hợp để tạo nên tất cả các tập thường xuyên kích thước $k + 1$, sau đó lại cắt xén, quét cơ sở dữ liệu... cho đến khi không tạo được tập thường xuyên. Trường hợp các tập đều là thường xuyên, thuật toán có độ phức tạp hàm mũ thì số lượng tính toán nhiều nhất. Tuy nhiên, thực tế dữ liệu mỏng đối với các loại bộ dữ liệu giao tác cần sử dụng thuật toán, bậc của tập thường xuyên lớn nhất thường là nhỏ, theo n . Khi đó thuật toán chỉ một lần quét tuyến tính cuối cùng, dùng danh sách các tập thường xuyên đã thấy. Nó xác định tổ hợp của các tập thường xuyên đáp ứng ngưỡng chính xác khi thể hiện luật, và cho ra luật kết hợp.

Khai phá dữ liệu quan tâm đến các thuật toán luật kết hợp ở chỗ việc tìm kiếm và tổ chức các thành phần quản trị dữ liệu là các thành phần chặt chẽ. Cụ thể, các thuật toán luật kết hợp sử dụng phương pháp tìm rộng, phương pháp tìm từ tổng quát đến cụ thể, cho phép tối thiểu số lần quét cơ sở dữ liệu. Trong khi có các thuật toán tìm khác trong học máy, sử dụng thể hiện luật, các thuật toán luật kết hợp được thiết kế đặc biệt cho cơ sở dữ liệu kích thước lớn một cách hiệu quả.

6.2.3.3. Thuật toán không gian vectơ trong tìm kiếm văn bản

Nhiệm vụ tổng quát của tìm theo nội dung được mô tả:

Định nghĩa: Tìm theo nội dung là khi có đối tượng hỏi và cơ sở dữ liệu lớn gồm các đối tượng, người ta cần tìm k đối tượng trong cơ sở dữ liệu giống đối tượng hỏi nhất.

Người ta quen thuộc với tìm trong văn bản trực tuyến. Ở đây xét thuật toán tìm tổng quát với các thành phần của nó. Tư liệu văn bản có độ dài, cấu trúc khác nhau. Ý tưởng chính để so sánh các tư liệu khác nhau là giảm các tư liệu về thể hiện vectơ.

Kí hiệu $t_1, \dots, t_p$ là p thuật ngữ, tức các từ, các câu... Có thể xem đó là các biến, cột trong ma trận dữ liệu. Một tư liệu, tức một dòng trong ma trận dữ liệu, được thể hiện bằng vectơ độ dài p , trong đó phần tử thứ i chứa số đếm mà thuật ngữ t_i xuất hiện trong tư liệu. Như bài toán về rổ hàng, người ta có cơ sở dữ liệu lớn, nhưng thưa. Dĩ nhiên người ta thường không lưu dữ liệu trong ma trận $n \times p$, mà thường dùng danh sách cho mỗi t_i của tất cả các tư liệu chứa t_i .

Cho thể hiện không gian vectơ, cách thể hiện độ giống nhau đơn giản là hàm khoảng cách tương tự, đo góc giữa hai vectơ trong không gian p chiều. Góc đo độ tương tự theo hướng đã cho theo thuật ngữ không gian, và nhân tố không khác nhau cho biết các tư liệu lớn có xu hướng xuất hiện nhiều lần một từ nào đó hơn các tư liệu nhỏ. Thể hiện không gian vectơ và độ đo tương tự theo góc có thể là đơn sơ, nhưng thực tế lược đồ này tồn tại được, và có một số biến thể.

Mô tả thuật toán phát hiện từ khóa (i) *nhiệm vụ*: tìm k tư liệu tương tự nhất trong cơ sở dữ liệu theo câu hỏi đã cho; (ii) *thể hiện*: xác định vectơ tần suất xuất hiện thuật ngữ; (iii) *hàm đích*: góc giữa hai vectơ; (iv) *phương pháp tìm*: các kĩ thuật khác nhau; (v) *kĩ thuật quản trị dữ liệu*: chiến lược chỉ số hóa nhanh đa dạng.

Có nhiều định nghĩa riêng đối với các thành phần trên. Chẳng hạn để xác định hàm đích, người ta có thể dùng độ đo tương tự

tổng quát hơn hàm góc. Về kĩ thuật tìm, lưu ý rằng tìm kiếm thời gian thực, do thuật toán cần tìm mẫu thời gian thực; không giống như các thuật toán khai phá dữ liệu trước cho phép tìm không trực tuyến.

Các ứng dụng khác nhau có thể dùng các thành phần trong thuật toán tìm. Người ta có thể xét ngữ cảnh ngược lại, tức xét các tư liệu cùng không chứa một số thuật ngữ.

Rõ ràng là thể hiện mô hình có vai trò quan trọng. Một khi dùng thể hiện vector, người ta có thể xác định dài rộng của các độ đo trong không gian vector, và có thể dùng kĩ thuật tìm, chỉ số hóa chuẩn để tìm điểm gần nhất trong không gian p chiều. Các thuật toán tìm khác nhau có thể thay đổi chút ít đối với hàm đích, nhưng đều chia sẻ việc thể hiện vector dữ liệu, sau khi thể hiện ở dạng vector. Tư liệu được phép dùng các thuật toán tìm.

6.2.4. Nhận xét chung về thuật toán khai phá dữ liệu

Đối với nhiều kĩ thuật, phương pháp, thuật toán dùng cho khai phá dữ liệu, cần có khuyến cáo áp dụng. Cách nhìn nhận theo thành phần nêu trong phần trên là công cụ hữu hiệu để ước lượng thuật toán. Quá trình gồm (i) cố loại bỏ phần khó hiểu không tránh được trong tài liệu nghiên cứu, sản phẩm; (ii) đưa thuật toán về dạng các thành phần cơ bản. Mô tả dựa trên thành phần cho phép hình dung khung cảnh để so sánh được, tức người ta có thể so sánh thuật toán mới với thuật toán nổi tiếng và nhìn chính xác về sự khác nhau theo nghĩa thành phần.

Việc nhấn mạnh sự khác nhau của thuật toán trong khai phá dữ liệu là có ý nghĩa trong cộng đồng xử lí dữ liệu. Về thống kê, cũng có nhiều phương trình xác định mô hình, hàm đích, phương pháp tính, và cả đặc tả chi tiết của thuật toán để mô hình khớp với thực tế. Về khoa học tính toán, trong học máy, nhận dạng người ta nhấn mạnh các phương pháp tính và thuật toán, và cả tính phù hợp của cấu trúc mô hình hay hàm đích cho các vấn đề nhấn mạnh. Khai phá dữ liệu tập trung vào hai lĩnh vực nghiên cứu đưa

đến các phát triển phương pháp luận hoàn toàn khác nhau. Các tiếp cận thống kê thường tập trung vào khía cạnh lý thuyết của các thủ tục suy luận, tức ước lượng tham số và chọn mô hình, và ít nhận vào tính toán. Tiếp cận khoa học tính toán đi ngược lại, tập trung vào cách tìm kiếm hiệu quả, quản trị dữ liệu, nhưng ít về sự phù hợp của mô hình, hay mẫu, hay sự hợp lý của hàm đích. Điều khác biệt này có giá trị, giúp giải thích các nhân tố kích hoạt sự phát triển của các mô hình riêng, các phương pháp suy luận, và các thuật toán. Cả hai trường phái thống kê, tính toán đều chưa làm rõ hết nhiều điều liên quan đến thành phần của thuật toán. Đối với bộ dữ liệu nhỏ, năng lực giải thích, dự đoán của mô hình là nhân tố quan trọng hơn khía cạnh tính toán. Khi bộ dữ liệu lớn, cả về số phép đo và số biến, vai trò của tính toán trở nên quan trọng nhiều. Ngoài ra, độ phức tạp về thời gian vẫn dùng giả thiết về dữ liệu trên bộ nhớ trong. Đối với dữ liệu nhiều, trên bộ nhớ ngoài, cần có kỹ thuật tổ chức dữ liệu để phù hợp với thuật toán; cần đến công nghệ thông tin nhiều hơn việc mô hình hóa thống kê truyền thống. Tất nhiên các cấu trúc mô hình và các hàm đích được chọn cẩn thận. Việc ước lượng cẩn thận giá tìm kiếm đối với mô hình, hay mẫu, lý tưởng và tài nguyên tính toán để đề xuất cấu trúc dữ liệu tin cậy.

6.3. Khai phá luật kết hợp

6.3.1. Nhu cầu về luật kết hợp

Trở lại với cơ sở dữ liệu về *Hàng điện tử*. Người ta muốn khuyên khách hàng mua hàng, dựa trên dữ liệu thống kê về thói quen mua hàng và mối liên quan giữa việc mua các mặt hàng. Khi đó cần đề cập (i) mẫu phổ biến; (ii) luật kết hợp, tức các tri thức đảm bảo phục vụ khách hàng.

Các mặt hàng, tức các đối tượng dữ liệu, được gọi là các mục hay *tập mục*¹. Trong hoạt động mua bán, người ta thực hiện các *giao dịch*¹, ứng với giao tác trong hệ thống cơ sở dữ liệu.

¹ Itemset: tập mục.

Định nghĩa: Mẫu phổ biến là mẫu, chẳng hạn tập mục, cấu trúc con, cấu trúc con, xuất hiện thường xuyên trong một bộ dữ liệu.

Các mặt hàng có thể hiện như bánh mì, máy ảnh, rau quả... thường xuất hiện trong giao dịch. Khi đó chúng là mẫu phổ biến. Trong một giao dịch mua bán, chúng xuất hiện lần lượt, tạo nên các *mục tuân tự*. Các cấu trúc đề cập là các dạng có cấu trúc (i) đồ thị con; (ii) cây con; (iii) dàn con, cho phép kết hợp các mục hay các dãy các mục. Nếu cấu trúc con xảy ra thường xuyên, nó được gọi là *mẫu cấu trúc phổ biến*. Việc phát hiện các mẫu phổ biến có ý nghĩa đối với (i) phát hiện mối quan hệ giữa các đối tượng; (ii) mối tương quan² trong dữ liệu; nên có ý nghĩa đối với nhiệm vụ khai phá dữ liệu.

Khai phá mẫu phổ biến tìm các mối quan hệ xảy ra trong bộ dữ liệu. Do vậy cần phát hiện (i) mối quan hệ; (ii) mối tương quan.

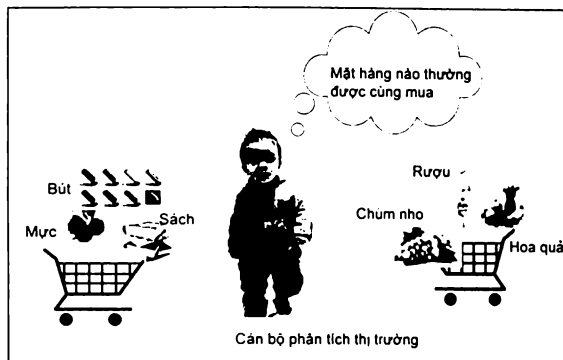
6.3.1.1. Phân tích mua hàng

Khai phá tập mục phổ biến dẫn đến sự phát hiện của các mối quan hệ và các mối tương quan giữa các mục trong bộ dữ liệu lớn quan hệ hay giao tác. Với số lượng lớn các dữ liệu, liên tục được cập nhật, người ta quan tâm nhiều đến việc khai phá mẫu như vậy từ cơ sở dữ liệu của họ. Kết quả được sử dụng cho (i) trợ giúp quyết định; (ii) phân tích hành vi khách hàng.

Thí dụ điển hình của khai phá tập mục phổ biến là *phân tích giỏ hàng*. Quá trình này phân tích thói quen mua sắm của khách hàng, thông qua tìm mối liên hệ giữa các mục khác nhau mà khách hàng đặt trong *giỏ hàng* của họ, tức trong giao dịch mua bán. Hãy xem xét một thí dụ về cách phân tích giỏ thị trường có thể có ích.

¹ Transaction: giao tác trong cơ sở dữ liệu, giao dịch trong hoạt động mua bán.

² Association: mối quan hệ; Correlation: mối tương quan.



Hình 6.5. Phân tích giỏ hàng

Về phân tích giỏ hàng: giả sử người quản lí của một chi nhánh Hàng điện tử muốn tìm hiểu thói quen mua sắm của khách hàng. Cụ thể, vấn đề là những nhóm hoặc bộ các mặt hàng mà khách hàng có phù hợp mua hàng trong lần đến các cửa hàng. Phân tích giỏ hàng được thực hiện trên các dữ liệu bán lẻ của các giao dịch của khách hàng tại cửa hàng. Người ta sử dụng kết quả phân tích để (i) hoạch định các chiến lược tiếp thị hay quảng cáo; hoặc (ii) thiết kế của một cửa hàng mới. Ngoài ra, các mục phổ biến được cùng mua có thể được đặt ở gần nhau, khuyến khích hơn việc bán kết hợp các mặt hàng đó.

Nếu xem không gian là tập hợp của các mặt hàng có sẵn tại cửa hàng, thì mỗi mục có một biến logic đại diện cho sự hiện diện hay vắng mặt của mục đó. Mỗi giỏ được biểu diễn bởi một vector logic của các giá trị được gán cho các biến này. Các vector logic được phân tích về các mẫu việc mua, phản ánh các mục, tức các mặt hàng, có (i) *quan hệ phổ biến*; hay (ii) cùng được mua. Những mẫu này có thể được biểu diễn dưới dạng *luật kết hợp*¹. Chẳng hạn thông tin về việc khách hàng mua màu vẽ cũng có khuynh hướng mua bút vẽ được thể hiện: *màu* $\Rightarrow$ *bút* (*trợ giúp* = 2%, *tin cậy* = 60%).

¹ Association rule: luật kết hợp.

Định nghĩa: Luật kết hợp là tri thức cho biết khả năng xảy ra của các mục, với độ trợ giúp và độ tin cậy.

Hai độ đo (i) trợ giúp luật; (ii) tin cậy là độ đo gắn với luật kết hợp. Trong thể hiện luật trên, trợ giúp 2% có nghĩa 2% các giao dịch theo phân tích cho rằng *màu* và *bút* được cùng mua; tin cậy 60% có nghĩa 60% khách hàng đã mua *màu* cũng đã mua *bút*. Người ta có khái niệm (i) *ngưỡng trợ giúp tối thiểu*; (ii) *ngưỡng tin cậy tối thiểu*. Người dùng có thể thiết lập các ngưỡng này. Ký hiệu $\Rightarrow$ dùng trong luật kết hợp, phân biệt với $\rightarrow$ trong (i) tri thức luật; hay (ii) logic kéo theo.

6.3.1.2. Tập mục phổ biến, tập mục đóng, luật kết hợp

Xét tập mục, tức tập các mặt hàng, $I = \{I_1, \dots, I_m\}$; dữ liệu thích hợp với nhiệm vụ D, là tập các giao tác, hay giao dịch mua bán, trong cơ sở dữ liệu. Đối tượng nghiên cứu ở đây là các *giao tác*, chứ không là các dữ liệu như cơ sở dữ liệu truyền thống. Giao dịch T được xác định như tập các mục, tức các mặt hàng, $T \subseteq I$. Mỗi giao dịch có tên, gọi là TID^1 .

Cho $A \subseteq I$, tức A gồm các mặt hàng. Giao dịch T chứa A nếu $A \subseteq T$. Luật kết hợp $A \Rightarrow B$ xác định với $A, B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ và $A \cap B = \emptyset$.

Định nghĩa: Trợ giúp s của luật kết hợp $A \Rightarrow B$, trong tập các giao dịch D, là tỉ lệ các giao dịch chứa A và B.

Giao dịch chứa A và B có nghĩa là chứa $A \cup B$. Trợ giúp luật cho biết khả năng xảy ra, xem như $P(A \cup B)$, tức $s(A \Rightarrow B) = P(A \cup B)$.

Định nghĩa: Tin cậy c của luật kết hợp $A \Rightarrow B$, trong tập các giao dịch D, là tỉ lệ các giao dịch đã có A, lại có B.

Do vậy, $c(A \Rightarrow B) = P(B|A)$.

¹ TID: Transaction Identifier: tên giao dịch.

Một số khái niệm khác:

- Các luật kết hợp thỏa mãn cả độ trợ giúp tối thiểu, *min-sup*, và độ tin cậy tối thiểu, *min-conf*¹, được gọi là *luật mạnh*;
- Toàn bộ tập mục có k mục được gọi là k -tập mục;
- *Tần suất xuất hiện của tập mục* là số các giao dịch chứa tập mục này. Người ta có thể gọi đơn giản là *tần suất*, *đếm trợ giúp*, hay *số đếm* của tập mục. Trợ giúp s trong $s(A \Rightarrow B) = P(A \cup B)$ là *trợ giúp tương đối*; tần suất xuất hiện là *trợ giúp tuyệt đối*;
- Tập mục I thỏa mãn *min-sup* xác định trước, tức trợ giúp tuyệt đối của I thỏa mãn ngưỡng số đếm trợ giúp tối thiểu, thì I là *tập mục phổ biến*.

Định nghĩa: Tập mục phổ biến là tập mục có tần suất xuất hiện thỏa mãn min-sup đã xác định.

Từ các khái niệm trên, người ta có:

$$c(A \Rightarrow B) = P(B | A) = \frac{s(A \cup B)}{s(A)} = \frac{\text{đếm trợ giúp}(A \cup B)}{\text{đếm trợ giúp}(A)}.$$

Công thức cho thấy độ tin cậy của các luật $A \Rightarrow B$ có thể dễ dàng rút ra từ số đếm trợ giúp của A và $A \cup B$. Tức khi có số trợ giúp của A , B , và $A \cup B$, dễ thu được mối quan hệ tương ứng với luật $A \Rightarrow B$ và $B \Rightarrow A$, và kiểm tra được luật mạnh. Do vậy, vấn đề của khai phá luật kết hợp đưa về việc khai phá tập mục phổ biến.

Nói chung, khai phá luật kết hợp có thể được xem như quá trình hai bước:

1. *Tìm tất cả các tập mục phổ biến.* Theo định nghĩa, mỗi tập mục phổ biến sẽ xảy ra phổ biến, ít nhất như *min-sup* đã xác định trước;

¹ Min – sup: minimal support threshold: ngưỡng trợ giúp tối thiểu; Min – conf: minimal confidence threshold: ngưỡng tin cậy tối thiểu.

2. *Tạo ra các luật kết hợp mạnh từ các tập mục phổ biến.* Những luật này đáp ứng trợ giúp tối thiểu và tin cậy tối thiểu.

Có thể sử dụng các độ đo khác, để phát hiện ra mối quan hệ, tương quan giữa các hạng mục. Bởi vì bước thứ hai là ít tốn kém hơn so với bước thứ nhất, hiệu suất tổng thể của khai phá luật kết hợp được xác định theo những bước đầu tiên.

Thách thức lớn trong khai phá tập mục phổ biến từ bộ dữ liệu lớn là việc khai phá như vậy thường tạo ra một số lượng lớn các tập mục phổ biến đáp ứng ngưỡng trợ giúp tối thiểu min-sup, đặc biệt khi min-sup nhỏ. Điều này là do nếu một tập là phổ biến thì tập con của nó cũng là phổ biến. Chẳng hạn tập mục phổ biến có độ dài 100 $\{a_1, \dots, a_{100}\}$; vậy sẽ có $\binom{100}{1} = 100$ tập 1 phổ biến $\{a_1\}, \dots, \{a_{100}\}$. Tổng các tập mục phổ biến là $\binom{100}{1} + \binom{100}{2} + \dots + \binom{100}{100} = 2^{100} - 1$. Đó là con số quá lớn. Để khắc phục điều này, người ta đưa ra khái niệm *tập mục phổ biến đóng* và *tập mục phổ biến cực đại*.

Định nghĩa: Tập mục X là đóng trong bộ dữ liệu D nếu không có tập trên¹ Y, mà Y có cùng số đếm trợ giúp như X trong D.

Tập mục X là một *tập mục phổ biến đóng* trong tập D nếu X là (i) đóng; (ii) phổ biến trong D.

Định nghĩa: Tập mục X là tập mục phổ biến cực đại, trong bộ dữ liệu D, nếu không có tập mục trên Y của X, mà Y là phổ biến.

Cho C là tập hợp các tập mục phổ biến đóng trong bộ dữ liệu D thỏa mãn ngưỡng trợ giúp min-sup. Gọi M là tập các tập mục phổ biến cực đại, trong D, thỏa mãn min-sup. Giả sử có số đếm trợ giúp của mỗi tập mục phổ biến trong C và M. Tập C và thông tin số đếm của nó có thể được sử dụng để rút ra toàn bộ các tập mục phổ biến.

¹ Super - itemset: tập mục trên; tập Y là tập trên của X nếu $X \subset Y$.

Như vậy, C có chứa đầy đủ thông tin liên quan đến các tập mục phổ biến của nó. Mặt khác, M chỉ ghi lại số trợ giúp của các tập mục phổ biến cực đại. Nó thường không có các thông tin trợ giúp đầy đủ về tập mục phổ biến của nó.

Về các tập mục phổ biến cực đại: giả sử cơ sở dữ liệu giao dịch chỉ có hai giao dịch: $\{(a_1, a_2, \dots, a_{100}); (a_1, \dots, a_{50})\}$. Lấy ngưỡng min-sup = 1. Có hai tập mục phổ biến $C = \{(a_1, a_2, \dots, a_{100}); (a_1, \dots, a_{50})\}$, và tập mục phổ biến cực đại $M = \{(a_1, a_2, \dots, a_{100})\}$. Tập các tập mục phổ biến đóng chứa đầy đủ thông tin liên quan đến các tập mục phổ biến. Từ C , người ta có thể thấy được tập $\{(a_i, a_i)\}$ là tập mục phổ biến, do nó là tập con của hai tập phổ biến kể trên.

6.3.2. Các phương pháp khai phá tập mục phổ biến

Phần này đề cập các phương pháp khai phá các dạng đơn giản nhất của mẫu phổ biến. Bắt đầu từ thuật toán Apriori, để tìm các tập mục phổ biến, phần này đề cập (i) cách tạo các luật kết hợp mạnh từ các tập mục phổ biến; (ii) biến thể của thuật toán Apriori để cải thiện hiệu suất và khả năng mở rộng; (iii) các phương pháp mẫu tăng trưởng để khai phá tập mục phổ biến, làm việc chỉ với không gian các tập mục phổ biến hiện tại; (iv) các phương pháp khai phá tập mục phổ biến, với định dạng dữ liệu theo chiều dọc.

6.3.2.1. Phát hiện tập mục phổ biến nhờ tạo ứng viên hạn chế

R. Agrawal và R. Srikant, năm 1994, đề xuất thuật toán Apriori¹ để khai phá các tập mục phổ biến cho luật kết hợp logic. Nó sử dụng tri thức tiên nghiệm về các tính chất của tập mục phổ biến. Apriori theo tiếp cận lặp, theo các mức; sử dụng k -tập mục để khai phá $(k + 1)$ -tập mục. (i) Trước tiên, phát hiện các 1-tập mục, nhờ duyệt cơ sở dữ liệu, tích lũy số đếm mỗi mục, rồi thu thập các mục thỏa mãn min-sup. Tập này là L_1 . (ii) Dùng L_1 để phát hiện các 2-tập mục. Việc này lặp đến khi không phát hiện thêm k -tập mục nào.

¹ Apriori: tên lấy theo *priori knowledge*, tức tri thức tiên nghiệm.

Việc tìm L_k đòi hỏi duyệt toàn bộ cơ sở dữ liệu. Để tăng hiệu quả của việc tạo theo mức các tập mục phổ biến, người ta sử dụng *tính chất Apriori*, để thu gọn không gian tìm kiếm.

Định nghĩa: Tính chất Apriori phát biểu rằng tất cả các tập con không rỗng của tập mục phổ biến cũng là phổ biến.

Tính chất này dựa trên những quan sát: tập mục I không là phổ biến, tức $P(I) < \min\text{-sup}$, thì khi bổ sung mục A vào I , tập mục $I \cup A$ vẫn không là phổ biến, tức $P(I \cup A) < \min\text{-sup}$. Tính chất theo cách *phản đơn điệu*¹, tức một tập không qua được thử nghiệm, thì tập trên của nó cũng không qua được, với cùng thử nghiệm đó.

Liên quan đến việc sử dụng tính chất Apriori trong thuật toán, người ta xem cách tạo L_k từ L_{k-1} . Người ta sử dụng quá trình hai bước (i) *nối*; (ii) *xén*².

- *Bước nối.* Để tìm L_k , cần tạo tập hợp các *ứng viên k* phổ biến, từ các tập $k-1$ phổ biến. Tập ứng viên k phổ biến được gọi là C_k . Lấy l_1, l_2 là tập mục trong L_{k-1} . Trong tập mục l_1 có các mục $l_1[j]$. Để thực hiện hiệu quả, Apriori giả định rằng các mục trong một giao dịch hoặc tập mục được sắp xếp theo thứ tự từ điển; tức $l_1[1] < l_1[2] \dots$. Phép toán L_{k-1} nối L_{k-1} được thực hiện nếu các phần tử của L_{k-1} có chung $k-2$ mục đầu tiên. Tức l_1 nối l_2 trong L_{k-1} nếu $l_1[1] = l_2[1], \dots, l_1[k-2] = l_2[k-2], l_1[k-1] < l_2[k-1]$. Kết quả của phép nối là tập mục $\{l_1[1], \dots, l_1[k-2], l_1[k-1], l_2[k-1]\}$;
- *Bước xen.* C_k là tập trên của L_k ; tức các thành viên của C_k có thể hoặc không phổ biến, nhưng tất cả các k tập mục phổ biến đều có trong C_k . Người ta duyệt cơ sở dữ liệu nhằm xác định số của mỗi ứng viên trong C_k ; việc này sẽ xác định L_k , tức tất cả các ứng viên với số đếm không ít hơn $\min\text{-sup}$ đều là phổ biến, và thuộc về L_k . Như vậy, để giảm

¹ Antimonotonicity: tính phản đơn điệu.

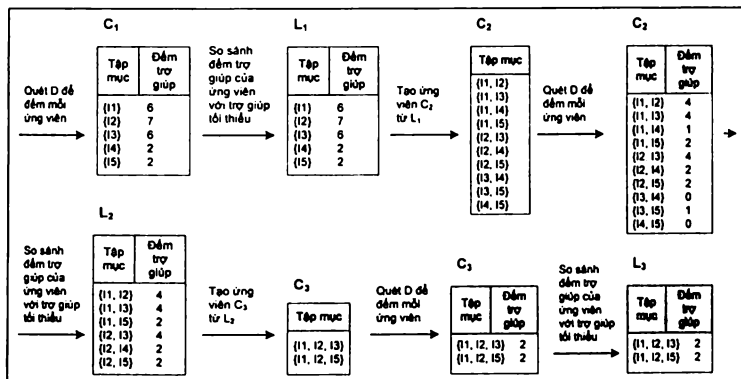
² Join, prune: nối, xen.

kích thước C_k , người ta sử dụng tính chất Apriori. Bất kì tập mục $k - 1$, không phổ biến, không thể là tập con của k tập mục. Vậy, nếu có tập con $k - 1$ trong ứng viên k tập mục không có trong L_{k-1} , thì ứng viên này không là phổ biến, và cần bị loại khỏi C_k . Việc *thử về tập con* được thực hiện nhanh, nhờ cây băm của tất cả các tập mục phổ biến.

Về Apriori, xét thí dụ cụ thể, dựa trên cơ sở dữ liệu giao dịch *Hàng điện tử D*, trong bảng sau. Có chín giao dịch trong cơ sở dữ liệu này, tức $|D| = 9$. Hình vẽ minh họa thuật toán tìm các tập mục phổ biến trong D .

Bảng 6.2. Dữ liệu giao dịch

Tên giao dịch TID	Danh sách các tên mục
T100	l_1, l_2, l_5
T200	l_2, l_4
T300	l_2, l_3
T400	l_1, l_2, l_4
T500	l_1, l_3
T600	l_2, l_3
T700	l_1, l_3
T800	l_1, l_2, l_3, l_5
T900	l_1, l_2, l_3



Hình 6.6. Tạo các tập mục ứng viên, với min-sup = 2

1. Trong lần lặp đầu của thuật toán, mỗi mục là một thành viên của tập hợp các ứng viên 1 - tập mục phổ biến, C_1 . Người ta duyệt đơn giản tất cả các giao dịch để đếm số lần xuất hiện của mỗi mục;
2. Giả sử $\text{min-sup} = 2$. Ở đây đang dùng trợ giúp *tuyệt đối*, vì đang sử dụng số đếm trợ giúp; trợ giúp *tương đối* là $2/9 = 22\%$. Tập các 1 tập mục phổ biến, L_1 , được xác định, chứa các 1 tập mục thỏa mãn min-sup . Ở đây các ứng viên C_1 đều thỏa mãn min-sup ;
3. Để phát hiện 2 tập mục, L_2 , thuật toán thực hiện L_1 nối L_1 , để tạo ứng viên cho 2 tập mục, C_2 . Tập C_2 có lực lượng là tổ hợp chập 2 của $|L_1|$. Không ứng viên nào trong C_2 bị loại trong bước xén, do mỗi tập con của nó đều là phổ biến;
4. Tiếp theo, các giao dịch trong D được duyệt và số trợ giúp của từng tập mục ứng viên trong C_2 được tích lũy;
5. Các 2 tập mục phổ biến, L_2 , được xác định, gồm những 2 tập mục phổ biến trong C_2 đáp ứng min-sup ;
6. Tạo tập các ứng viên 3 tập mục, C_3 . $C_3 = L_2$ nối $L_2 = \{\{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_5\}, \{I_2, I_3\}, \{I_2, I_4\}, \{I_2, I_5\}\}$ nối $\{\{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_5\}, \{I_2, I_3\}, \{I_2, I_4\}, \{I_2, I_5\}\} = \{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}, \{I_1, I_3, I_5\}, \{I_2, I_3, I_4\}, \{I_2, I_3, I_5\}, \{I_2, I_4, I_5\}\}$. Bước xén sử dụng tính chất Apriori, tức tất cả các tập con của tập mục phổ biến cần là phổ biến. Người ta xét các tập con: (i) các 2 tập mục con của $\{I_1, I_2, I_3\}$ là $\{I_1, I_2\}, \{I_1, I_3\}, \{I_2, I_3\}$, đều là phần tử của L_2 . Do vậy người ta giữ lại $\{I_1, I_2, I_3\}$ trong C_3 . (ii) Các 2 tập mục của $\{I_1, I_2, I_5\}$ là $\{I_1, I_2\}, \{I_1, I_5\}, \{I_2, I_5\}$, đều là phần tử của L_2 . Do vậy tập mục này được giữ trong C_3 . (iii) Các 2 tập mục của $\{I_1, I_3, I_5\}$ là $\{I_1, I_3\}, \{I_3, I_5\}, \{I_1, I_5\}$. Nhận xét thấy $\{I_3, I_5\}$ không là phần tử của L_2 , nên không là phổ biến. Vậy $\{I_1, I_3, I_5\}$ bị loại khỏi C_3 . (iv) Các 2 tập mục của $\{I_2, I_3, I_4\}$ là $\{I_2, I_3\}, \{I_3, I_4\}, \{I_2, I_4\}$. Nhận xét thấy $\{I_3, I_4\}$ không là phần tử của L_2 , nên không là phổ biến. Vậy $\{I_2, I_3, I_4\}$ bị loại khỏi C_3 . (v) Các 2 tập mục

của $\{I_2, I_3, I_5\}$ là $\{I_2, I_3\}$, $\{I_3, I_5\}$, $\{I_2, I_5\}$. Nhận xét thấy $\{I_3, I_5\}$ không là phần tử của L_2 , nên không là phổ biến. Vậy $\{I_2, I_3, I_5\}$ bị loại khỏi C_3 . (vi) Các 2 tập mục của $\{I_2, I_4, I_5\}$ là $\{I_2, I_4\}$, $\{I_4, I_5\}$, $\{I_2, I_5\}$. Nhận xét thấy $\{I_4, I_5\}$ không là phần tử của L_2 , nên không là phổ biến. Vậy $\{I_2, I_4, I_5\}$ bị loại khỏi C_3 . Do vậy $C_3 = \{\{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}\}$;

7. Các giao dịch trong D được duyệt để xác định L_3 , gồm những ứng viên 3 tập mục phổ biến trong C_3 , thỏa mãn min-sup;
8. Thuật toán thực hiện L_3 nối L_3 để tạo ra ứng viên của 4 tập mục phổ biến, C_4 . Người ta được $\{\{I_1, I_2, I_3, I_5\}\}$, nhưng tập con $\{I_2, I_3, I_5\}$ của nó không là phổ biến, nên nó bị loại. Vậy $C_4 = \emptyset$; thuật toán kết thúc với các tập mục phổ biến đã có.

Tiếp theo là thể hiện giả mã của thuật toán Apriori và các thủ tục liên quan.

Thuật toán Apriori phát hiện các tập mục phổ biến, dùng tiếp cận lặp theo các mức, bằng cách tạo các ứng viên.

Đầu vào: (i) tập các giao dịch D ; (ii) ngưỡng trợ giúp tối thiểu, min-sup.

Đầu ra: Các tập mục L trong D .

Phương pháp:

L_1 = phát hiện các 1 tập mục phổ biến trong D ;

For ($k = 2; L_{k-1} \neq \emptyset; k++$) {

C_k = tạo Apriori (L_{k-1});

For each giao dịch $t \in D$ { // duyệt D để tìm số đếm

C_t = tập con (C_k, t); // lấy các tập con ứng

viên

For each ứng viên $c \in C_t$

$c.\text{đếm}++$;

}


```

    }
    Return L = {Lk};

Procedure tạo Apriori (Lk-1: k - 1 tập mục phổ biến)
For each tập mục l1 ∈ Lk-1
    For each tập mục l2 ∈ Lk-1
        If (l1[1] = l2[1], ..., l1[k - 2] = l2[k - 2],
        l1[k - 1] < l2[k - 1] l1[k - 2] = l2[k - 2], l1[k - 1] <
        l2[k - 1]) then {
            c = l1 nối l2; // bước nối, tạo các ứng viên
            If có tập không phổ biến (c, Lk = 1)
then
            Xóa s; // bước xén, loại bỏ ứng viên không phù hợp
        Else bổ sung c vào Ck;
    }
Return Ck;

Procedure có tập không phổ biến (c: ứng viên k tập
mục; Lk-1: (k - 1) tập mục phổ biến) // sử dụng tri thức
Apriori
    For each (k - 1) tập mục s của c
        If s ∉ Lk-1 then
            Return True;
    Return False;

```

6.3.2.2. Tạo luật kết hợp từ tập mục phổ biến

Một khi đã có các tập mục phổ biến từ các giao dịch trong cơ sở dữ liệu D, việc tìm các luật kết hợp mạnh là đơn giản. Luật kết hợp mạnh thỏa mãn cả trợ giúp tối thiểu và tin cậy tối thiểu:

$$c(A \Rightarrow B) = P(B | A) = \frac{s(A \cup B)}{s(A)} = \frac{\text{đem tro giúp}(A \cup B)}{\text{đem tro giúp}(A)}$$

Xác suất có điều kiện được thể hiện dưới dạng số đếm trợ giúp, trong đó số đếm ($A \cup B$) là số các giao dịch chứa các tập mục $A \cup B$, và số đếm trợ giúp (A) là số các giao dịch chứa tập mục A . Các luật kết hợp được tạo:

1. Đối với mỗi tập mục phổ biến l phổ biến, tạo ra tất cả các tập con khác rỗng của l ;
2. Đối với mỗi tập con s khác rỗng của l , sinh ra luật $s \Rightarrow (l - s)$ nếu đếm trợ giúp (l) / đếm trợ giúp (s) $\geq$ min-conf; trong đó min-conf là ngưỡng tin cậy tối thiểu.

Bởi vì các luật được tạo ra từ các tập mục phổ biến, mỗi luật đã thỏa mãn min-sup. Các tập mục phổ biến được lưu trong bảng băm, theo số đếm của nó, để dễ truy cập.

Vẽ tạo luật kết hợp: quay lại thí dụ các giao dịch, đã liệt kê trong bảng phân trước. Dữ liệu này có tập mục phổ biến $X = \{I_1, I_2, I_5\}$. Người ta tạo các luật kết hợp từ X . Các tập con khác rỗng của X là $\{I_1, I_2\}$, $\{I_1, I_5\}$, $\{I_2, I_5\}$, $\{I_1\}$, $\{I_2\}$, $\{I_5\}$. Các luật là:

$\{I_1, I_2\} \Rightarrow \{I_5\}$, với tin cậy = $2 / 4 = 50\%$;

$\{I_1, I_5\} \Rightarrow \{I_2\}$, với tin cậy = $2 / 2 = 100\%$;

$\{I_2, I_5\} \Rightarrow \{I_1\}$, với tin cậy = $2 / 2 = 100\%$;

$\{I_1\} \Rightarrow \{I_2, I_5\}$, với tin cậy = $2 / 6 = 33\%$;

$\{I_2\} \Rightarrow \{I_1, I_5\}$, với tin cậy = $2 / 7 = 29\%$;

$\{I_5\} \Rightarrow \{I_1, I_2\}$, với tin cậy = $2 / 2 = 100\%$.

Nếu min-conf = 70%, người ta được các luật thứ hai, ba và sáu. Không giống như các luật phân lớp thông thường, luật kết hợp có thể chứa nhiều hơn một liên kết ở phía bên phải của luật.

6.3.2.3. Nâng cao hiệu quả của Apriori

Để nâng cao hiệu quả khai phá theo thuật toán Apriori, người ta có nhiều biến thể của Apriori. Một số các biến thể được tóm tắt như sau:

- *Kĩ thuật dựa trên băm*. Kĩ thuật này băm các tập mục thành các gói. Nó cho phép giảm kích thước của các ứng viên k tập mục phổ biến, C_k , $k > 1$. Chẳng hạn khi duyệt mỗi giao dịch trong cơ sở dữ liệu để tạo các 1 tập mục phổ biến, L_1 , có thể tạo ra tất cả 2 tập mục phổ biến đối với mỗi giao dịch, băm, tức khớp, chúng vào các gói khác nhau của cấu trúc *bảng băm*, và tăng số đếm của gói. Một 2 tập mục với số thùng tương ứng trong bảng băm mà dưới ngưỡng trợ giúp không thể là phổ biến, và cần bị loại bỏ từ tập ứng viên. Kĩ thuật dựa trên băm này có thể làm giảm đáng kể số các ứng viên k tập mục phổ biến xem xét, đặc biệt khi $k = 2$;

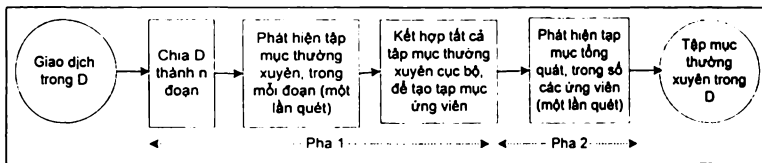
H₂							
Địa chỉ bó	0	1	2	3	4	5	6
Số đếm bó	2	2	4	2	2	4	4
Nội dung bó	{11, 14} {13, 15}	{11, 15} {11, 15}	{12, 13} {12, 13} {12, 13} {12, 13}	{12, 14} {12, 14}	{12, 15} {12, 15}	{11, 12} {11, 12} {11, 12}	{11, 13} {11, 13} {11, 13} {11, 13}

Tạo bảng băm H_2 ,
theo hàm băm $h(x, y) = ((\text{thứ tự của } x) * 10 + (\text{thứ tự của } y) * 7) \bmod 7$

Hình 6.8. Bảng băm H_2 , đối với ứng viên 2 tập mục phổ biến

- *Giảm giao dịch*. Việc này giảm số duyệt các giao dịch trong lần lặp sắp tới. Một giao dịch không chứa bất kỳ k tập mục phổ biến sẽ không thể chứa bất kỳ $(k + 1)$ tập mục phổ biến. Do đó có thể đánh dấu hay loại bỏ nó trong lần xem xét sau;
- *Phân hoạch*. Nó phân hoạch dữ liệu để phát hiện các tập mục ứng viên. Một kĩ thuật phân hoạch có thể sử dụng chỉ cần hai lần duyệt cơ sở dữ liệu, để khai phá tập mục phổ biến. Có hai pha. (i) Pha đầu chia các giao dịch trong D thành n đoạn không chồng nhau. Nếu ngưỡng trợ giúp tương đối tối thiểu đối với các giao dịch trong D là $\min\text{-sup}$,

thì số đếm trợ giúp tối thiểu đối với đoạn này là $\min\text{-sup}^*$ số các giao dịch trong đoạn này. Đối với mỗi đoạn, phát hiện được tất cả các tập mục phổ biến địa phương, tức các tập mục phổ biến trong đoạn này. Một tập mục phổ biến địa phương có thể hoặc không là trong toàn D . Bất kỳ tập mục phổ biến trong D , cần là tập mục phổ biến trong ít nhất trong một đoạn. Do vậy, tất cả các tập mục phổ biến địa phương là tập mục phổ biến ứng viên trong D . Tập các tập mục tập mục phổ biến từ tất cả các đoạn tạo nên các tập mục phổ biến ứng viên toàn thể, đối với D . (ii) Pha thứ hai thực hiện duyệt cơ sở dữ liệu lần hai, đánh giá các tập mục phổ biến toàn thể, theo các trợ giúp thực. Kích thước của đoạn và số các đoạn được thiết lập phù hợp với hạ tầng kỹ thuật;



Hình 6.9. Khai phá nhờ phân đoạn dữ liệu

- Lấy mẫu.** Tiếp cận này khai phá trên tập con của bộ dữ liệu D . Ý tưởng cơ bản của phương pháp lấy mẫu là chọn ngẫu nhiên mẫu S bộ dữ liệu D , rồi tìm kiếm các tập mục phổ biến trong S thay vì D . Người ta trả giá cho việc thu gọn dữ liệu bằng độ chính xác. Kích thước S mẫu được chọn để khai phá tại bộ nhớ chính, và duyệt toàn thể các giao dịch. Do tìm kiếm các tập mục phổ biến trong S thay cho D , có thể sót một số các tập mục phổ biến toàn thể. Để giảm trường hợp như vậy, người ta dùng ngưỡng trợ giúp thấp hơn $\min\text{-sup}$, để phát hiện các tập mục phổ biến địa phương tại, kí hiệu L^S . Phần còn lại của cơ sở dữ liệu sau đó được sử dụng để tính tần số thực của mỗi tập mục phổ biến trong L^S . Một cơ chế được sử dụng để xác định xem tất cả các tập mục phổ biến toàn thể có trong L^S hay không.

Nếu L^S thực sự chứa tất cả các tập mục phổ biến trong D , người ta chỉ cần duyệt D một lần. Nếu không, người ta duyệt lần nữa, để phát hiện các tập mục phổ biến mà đã bị bỏ sót trong lần duyệt đầu. Phương pháp lấy mẫu là đặc biệt có lợi khi hiệu quả là vô cùng quan trọng, khi thường xuyên chạy các ứng dụng tính toán chuyên sâu;

- *Đếm tập mục động.* Việc này bổ sung các tập mục ứng viên tại các điểm khác nhau trong lúc duyệt. Kỹ thuật đếm tập mục phổ biến động chia cơ sở dữ liệu thành các khối, đánh dấu bởi các điểm bắt đầu. Các tập mục ứng viên mới có thể được bổ sung tại bất kỳ điểm bắt đầu, không như trong Apriori, mà quyết định tập mục ứng viên mới chỉ ngay trước mỗi lần duyệt toàn bộ cơ sở dữ liệu. Kỹ thuật này sử dụng *số đếm*¹ như biên thấp của số đếm thực sự. Nếu *số đếm* vượt quá min-sup, tập mục này được thêm vào bộ sưu tập các tập mục phổ biến và có thể được sử dụng để tạo ra các ứng viên dài hơn. Điều này cho phép duyệt cơ sở dữ liệu ít đi, so với Apriori, khi phát hiện tất cả các tập mục phổ biến.

6.3.2.4. Tiếp cận tăng trưởng mẫu để khai phá tập mục phổ biến

Trong nhiều trường hợp, phương pháp *tạo và thử*² các ứng viên trong thuật toán Apriori giảm đáng kể kích thước của tập các ứng viên, tức tăng hiệu suất. Tuy nhiên, nó phải chịu hai chi phí đáng kể sau:

1. *Nó vẫn có thể cần phải tạo ra một số lượng lớn các ứng viên.* Chẳng hạn, nếu có 10^4 1 tập mục phổ biến, thuật toán Apriori sẽ cần phải tạo ra hơn 10^7 ứng viên 2 tập mục phổ biến;
2. *Nó có thể cần phải lặp lại việc duyệt toàn bộ cơ sở dữ liệu và kiểm tra một tập hợp lớn các ứng viên để khớp mẫu.* Việc duyệt

¹ Count - so - far: số đếm, trong kỹ thuật đếm động.

² Generate - and - test: tạo và thử.

mỗi giao dịch trong cơ sở dữ liệu để xác định số trợ giúp của các tập mục ứng viên là chi phí đáng kể.

Người ta đặt vấn đề về khả năng thiết kế một phương pháp mà khai phá tập đầy đủ các tập mục phổ biến, mà không sử dụng quá trình tạo các ứng viên tốn kém. Có phương pháp với tên *tăng trưởng mẫu phổ biến*, hay *tăng trưởng FP*¹, sử dụng chiến lược *chia để trị*². (i) Trước tiên, cơ sở dữ liệu thể hiện các mục phổ biến bị ép thành *cây mẫu phổ biến*, hay *cây FP*, mà vẫn giữ được thông tin liên kết tập mục. Rồi phân chia cơ sở dữ liệu này vào tập các *cơ sở dữ liệu điều kiện*, tức một loại đặc biệt do phép chiếu cơ sở dữ liệu. Mỗi cơ sở dữ liệu liên kết với một mục phổ biến hoặc *đoạn mẫu*, và khai phá mỗi cơ sở dữ liệu theo cách tách biệt. Đối với mỗi đoạn mẫu, chỉ xét bộ dữ liệu liên quan. Vì vậy, phương pháp này có thể giảm đáng kể kích thước của các bộ dữ liệu được tìm kiếm, cùng với sự *tăng trưởng* của mẫu được xét.

Về tăng trưởng FP: phát hiện tập mục phổ biến mà không cần tạo ứng viên. Xét việc khai phá cơ sở dữ liệu giao dịch D đã đề cập.

Lần duyệt đầu tiên, như Priori, rút ra tập các tập mục phổ biến, 1 tập mục, và số đến trợ giúp, tức tần suất. Lấy $\min\text{-sup} = 2$. Người ta sắp xếp các mục phổ biến, theo thứ tự giảm của số đếm, tạo nên tập, hay *danh sách*, L. Người ta có $L = \{\{I_2:7\}, \{I_1:6\}, \{I_3:6\}, \{I_4:2\}, \{I_5:2\}\}$.

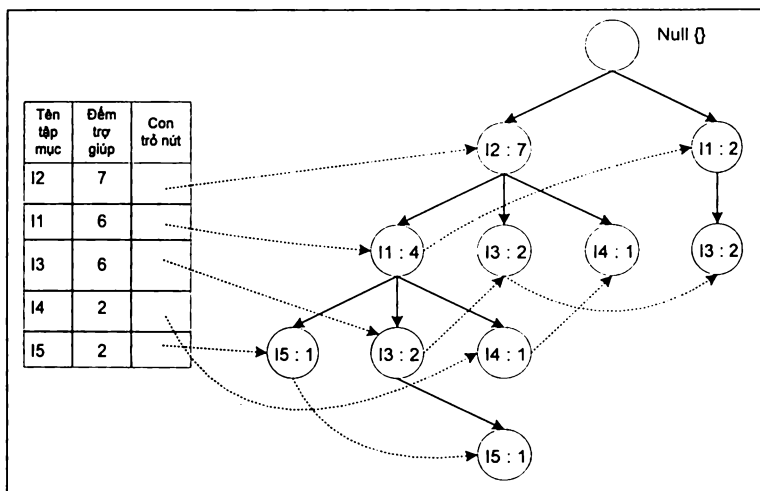
Cây FP được xây dựng như sau. Đầu tiên, tạo thư mục gốc của cây, dán nhãn *null*. Quét cơ sở dữ liệu D lần thứ hai. Các mục trong mỗi giao dịch được xử lý theo thứ tự L, tức đã sắp xếp giảm dần về số trợ giúp, và nhánh cây được tạo ra cho mỗi giao dịch. Chẳng hạn, (i) quá trình duyệt giao dịch đầu tiên, *T100*: I_1, I_2, I_5 , trong đó có ba mục I_2, I_1, I_5 trong L theo thứ tự, cho phép xây dựng nhánh đầu tiên của cây với ba nút, ($I_2: 1$), ($I_1: 1$), và ($I_5: 1$). Nút I_2 được liên kết như con thư mục gốc, I_1 liên kết với I_2 , và I_5 được liên

¹ FP - growth: frequent pattern growth: tăng trưởng mẫu phổ biến.

² Divide - and - conquer: chia để trị.

kết với I_1 . (ii) Xét giao dịch thứ hai, $T200:I_2:I_4$, chứa các mục I_2 và I_4 theo thứ tự trong L , cho phép tạo nhánh cây I_2 được liên kết vào thư mục gốc và I_4 được liên kết với I_2 . Tuy nhiên, nhánh I_2 này sẽ chia sẻ *tiền tố* chung, I_2 , của nhánh ứng với $T100$. Do đó, người ta tăng số đếm I_2 , và tạo nút con I_4 .

Để dễ duyệt cây, người ta xây dựng *bảng các nút*, để mỗi mục trở đến các thể hiện của mục trên cây, tạo nên dãy các con trỏ, tức *liên kết các nút*. Khi đó việc khai phá các mẫu phổ biến trong cơ sở dữ liệu chuyển về khai phá cây FP.



Hình 6.10. Cây FP mang thông tin về mẫu phổ biến

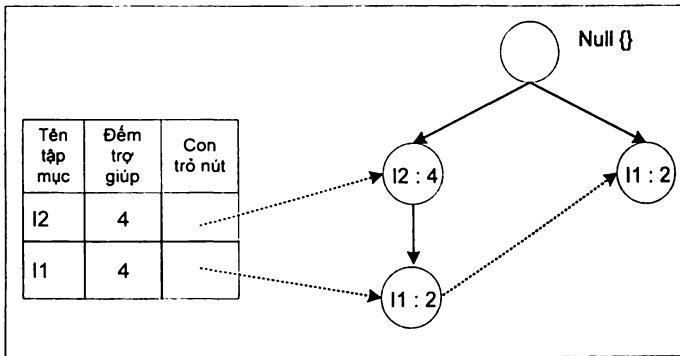
Khai phá cây FP bắt đầu từ mỗi mẫu độ dài 1, xem như *hậu tố của mẫu*, người ta kiến thiết *cơ sở mẫu điều kiện*, tức cơ sở dữ liệu con gồm tập các *đường dẫn tiền tố* trong cây FP, cùng xảy ra với hậu tố. Rồi kiến thiết cây FP, có điều kiện, và thực hiện khai phá đệ quy trên cây. Sự phát triển mẫu hình thành bằng cách nối của mẫu hậu tố với các mẫu phổ biến được tạo ra từ cây FP điều kiện.

Việc khai phá cây FP được tóm tắt như trong bảng. Trước tiên xét I_5 , là mục cuối cùng trong L . Người ta xuất phát từ cuối danh sách do như trong hình, để phát hiện nút I_5 cuối cây FP; có các

nhánh $(I_2, I_1, I_5:1^1)$ và $(I_2, I_1, I_3, I_5:1)$. Xem I_5 là hậu tố, với hai nhánh tiền tố là $(I_2, I_1:1^2)$ và $(I_2, I_1, I_3:1)$, tạo nên cơ sở mẫu điều kiện. Sử dụng cơ sở mẫu điều kiện này như là một cơ sở dữ liệu giao dịch, người ta xây dựng cây FP điều kiện I_5 , chỉ chứa một nhánh $(I_2, I_1:1)$; I_3 không có mặt, do có trợ giúp là 1, nhỏ hơn min-sup, 2. Nhánh này tạo nên tất cả các tổ hợp của mẫu phổ biến $\{I_2, I_5:2\}$, $\{I_1, I_5:2\}$, $\{I_2, I_1, I_5:2\}$.

Bảng 6.3. Khai phá cây FP nhờ tạo cơ sở mẫu điều kiện

Mục	Cơ sở mẫu điều kiện	Cây FP điều kiện	Các mẫu phổ biến được tạo
I_5	$\{\{I_2, I_1:1\}, \{I_2, I_1, I_3:1\}\}$	$(I_2:2, I_1:2)$	$\{I_2, I_5:2\}$, $\{I_1, I_5:2\}$, $\{I_2, I_1, I_5:2\}$
I_4	$\{\{I_2, I_1:1\}, \{I_2:1\}\}$	$(I_2:2)$	$\{I_2, I_4:2\}$
I_3	$\{\{I_2, I_1:2\}, \{I_2:2\}, \{I_1:2\}\}$	$(I_2:4, I_1:2), (I_1:2)$	$\{I_2, I_3:4\}$, $\{I_1, I_3:4\}$, $\{I_2, I_1, I_3:2\}$
I_1	$\{I_2:4\}$	$(I_2:4)$	$\{I_2, I_1:4\}$



Hình 6.11. Cây FP điều kiện liên kết với nút điều kiện I_3

Thuật toán tăng trưởng FP, khai phá các tập mục phổ biến, dùng cây FP, nhờ tăng trưởng đoạn mẫu.

Đầu vào: (i) cơ sở dữ liệu giao dịch D ; (ii) ngưỡng đếm trợ giúp tối thiểu min-sup.

¹ Riêng I_5 có $I_5:1$, để nhận biết phần tử rẽ nhánh.

² Lúc này I_1 có trợ giúp 1, nên ghi là $I_1:1$.

Đầu ra: tập đầy đủ của các đoạn phổ biến.

Phương pháp:

Cây FP được kiến thiết theo các bước sau:

1. Duyệt cơ sở dữ liệu một lần. Thu thập tập các mục phổ biến F , và số đếm trợ giúp. Sắp xếp F theo thứ tự giảm của số đếm, tạo được danh sách L .
2. Tạo gốc cây FP, gán nhãn null. Đối với mỗi giao dịch T trong D , thực hiện: Chọn và sắp xếp các mục trong T theo thứ tự của L . Lấy danh sách các mục phổ biến trong T là $[p|P]$, trong đó p là phần tử đầu tiên và P là danh sách còn lại. Thực hiện *bổ sung cây* ($[p|P]$, T), như sau. Nếu T có con N mà $N.tên = p.tên$ thì tăng số đếm của N thêm 1; ngược lại tạo nút N mới, và lấy số đếm là 1, cho con trỏ cha của nó trỏ đến T , và cho liên kết nút của nó trỏ đến các nút với cùng *tên*, thông qua cấu trúc liên kết nút. Nếu P là khác rỗng, thực hiện *đề quy bổ sung cây* (P , N).

Khai phá cây FP nhờ thực hiện *tăng trưởng FP* (cây FP, null).

Procedure tăng trưởng $FP(T, a)$ // T là cây

If T chứa đường P đơn then

For each tổ hợp, kí hiệu b , của các nút trong đường P

Tạo mẫu $b \cup a$, với đến trợ giúp = đếm trợ giúp tối thiểu của các nút trong b ;

Else for each a_i trong phần đầu của T {

Tạo mẫu $b = a_i \cup a$ với đếm trợ giúp = a_i .đếm trợ giúp;

Kiến thiết cơ sở mẫu dữ liệu của b và cây FP điều kiện của b , T_b ;

If $T_b \neq \emptyset$ then

Thực hiện *tăng trưởng FP* (T_b , b);

Thuật toán tăng trưởng FP để khai phá các tập mục phổ biến, không tạo ứng viên.

Phương pháp tăng trưởng FP chuyển vấn đề phát hiện mẫu phổ biến dài thành tìm kiếm đệ quy ngắn hơn trong cơ sở dữ liệu điều kiện nhỏ hơn nhiều, và rồi nối các hậu tố. Nó sử dụng các mục phổ biến cuối làm hậu tố. Phương pháp này làm giảm đáng kể chi phí tìm kiếm.

Khi cơ sở dữ liệu lớn, đôi khi xây dựng cây FP trong bộ nhớ là không thực tế. Người ta sử dụng thay thế: (i) chia cơ sở dữ liệu thành tập các cơ sở dữ liệu nhờ phép chiếu; (ii) kiến thiết cây FP; (iii) khai phá cây trong mỗi cơ sở dữ liệu chiếu. Quá trình này có thể dùng đệ quy đối với bất kì cơ sở dữ liệu chiếu mà kích thước chưa phù hợp với bộ nhớ trong.

6.3.2.5. Khai phá tập mục phổ biến nhờ dạng thức dữ liệu dọc

Cả hai phương pháp Apriori và tăng trưởng FP đều khai phá mẫu phổ biến từ tập các giao dịch theo định dạng TID của tập mục, tức $\{TID: \text{tập mục}\}$, trong đó TID là một tên giao dịch và *tập mục* là tập các mục mua trong giao dịch TID . Dạng này được xem như *định dạng dữ liệu ngang*. Ngoài ra, dữ liệu có thể thể hiện dưới dạng *mục - tập các TID*, tức $\{mục: \text{tập các TID}\}$, trong đó *mục* là tên mục giao dịch, và *tập các TID* là tập các tên giao dịch chứa các mục. Đây là *dạng dữ liệu dọc*¹.

Phần này đề cập cách các tập mục phổ biến có thể được khai phá hiệu quả với dạng dữ liệu dọc, với thuật toán chuyển lớp tương đương Eclat².

Về khai phá các tập mục phổ biến với dạng dữ liệu dọc: xét các định dạng dữ liệu theo chiều ngang của cơ sở dữ liệu giao dịch, D , trong các thí dụ trước. Chúng có thể được chuyển thành dạng dữ liệu dọc, nhờ duyệt một lần cơ sở dữ liệu.

¹ Vertical data format: dạng dữ liệu dọc; Horizontal data format: dạng dữ liệu ngang.

² Eclat Equivalence Class Transformation: chuyển lớp tương đương.

Bảng 6.4. Dạng dữ liệu dọc do chuyển dữ liệu từ thí dụ trước

Mục	Tập TID
I_1	{T100, T400, T500, T700, T800, T900}
I_2	{T100, T200, T300, T400, T600, T800, T900}
I_3	{T300, T500, T600, T700, T800, T900}
I_4	{T200, T400}
I_5	{T100, T800}

Khai phá có thể thực hiện trên dữ liệu này, bằng cách giao các tập TID của mọi mục phổ biến đơn. Lấy min-sup = 2. Do mỗi mục đơn là phổ biến trong bảng trên, có 10 phép giao, tạo nên tám 2-tập mục không rỗng, trong bảng tiếp theo. Do các tập mục $\{I_1, I_4\}$ và $\{I_3, I_5\}$ chỉ gồm một giao dịch, nên chúng không thuộc vào tập các 2-tập mục phổ biến.

Bảng 6.5. Các 2-tập mục, do phép giao dữ liệu bảng trên

Tập mục	Tập TID
$\{I_1, I_2\}$	{T100, T400, T800, T900}
$\{I_1, I_3\}$	{T500, T700, T800, T900}
$\{I_1, I_4\}$	{T400}
$\{I_1, I_5\}$	{T100, T800}
$\{I_2, I_3\}$	{T300, T600, T800, T900}
$\{I_2, I_4\}$	{T200, T400}
$\{I_2, I_5\}$	{T100, T800}
$\{I_3, I_5\}$	{T800}

Dựa theo tính chất Apriori, 3-tập mục là 3-tập mục ứng viên chỉ khi mọi tập con của 2-tập mục là phổ biến. Ứng viên cho 3-tập mục là $\{I_1, I_2, I_3\}$ và $\{I_1, I_2, I_5\}$. Thực hiện giao các tập TID của hai 2-tập mục của các ứng viên 3-tập mục, được kết quả như bảng sau.

Bảng 6.6. Các 3-tập mục theo dạng thức dữ liệu dọc

Tập mục	Tập TID
$\{I_1, I_2, I_3\}$	{T800, T900}
$\{I_1, I_2, I_5\}$	{T100, T800}

Thí dụ trên đã minh họa quá trình khai phá các tập mục phổ biến bằng cách khai thác dạng dữ liệu dọc. Đầu tiên, người ta

chuyển dữ liệu định dạng ngang sang định dạng dọc, bằng cách duyệt bộ dữ liệu một lần. Số đếm trợ giúp của tập mục phổ biến chỉ đơn giản là chiều dài của tập TID của tập mục phổ biến. Bắt đầu với $k = 1$, k -tập mục phổ biến có thể được sử dụng để xây dựng các ứng viên $(k + 1)$ -tập mục dựa trên tính chất Apriori.

Việc tính toán được thực hiện nhờ lấy giao của các bộ TID của k -tập mục phổ biến, để tính các bộ TID của $(k + 1)$ -tập mục tương ứng. Quá trình này lặp đi lặp lại, với k tăng thêm 1 mỗi lần, cho đến khi không thấy tập mục phổ biến hoặc ứng viên tập mục phổ biến.

Bên cạnh ưu điểm lợi dụng tính chất Apriori trong lúc tạo ứng viên $(k + 1)$ -tập mục từ k -tập mục phổ biến, một ưu điểm của phương pháp này là không cần duyệt cơ sở dữ liệu để tìm trợ giúp của $(k + 1)$ -tập mục, $k \geq 1$. Sờ dĩ như vậy là vì tập TID của mỗi k -tập mục mang đầy đủ thông tin cần thiết để tính trợ giúp này. Tuy nhiên, tập TID có thể khá dài, cần không gian nhớ đáng kể, cũng như thời gian tính toán khi giao các tập dài.

Để giảm hơn nữa chi phí đối với các tập TID dài, cũng như chi phí của phép giao, người ta có thể sử dụng kỹ thuật, cho phép *theo vết chỉ sự khác nhau* của các tập TID của một $(k + 1)$ -tập mục và k -tập mục tương ứng. Chẳng hạn trong thí dụ trên, $\{I_1\} = \{T100, 1400, 1500, T700, T800, T900\}$ và $\{I_1, I_2\} = \{T100, T400, T800, T900\}$. Sai khác giữa hai tập là *sai khác* $(\{I_1, I_2\}, \{I_1\}) = \{T500, T700\}$. Như vậy, thay vì ghi bốn TIDs tạo nên giao của $\{I_1\}$ và $\{I_2\}$, người ta sử dụng sai khác để ghi hai TID, cho biết sự khác nhau giữa $\{I_1\}$ và $\{I_1, I_2\}$. Kinh nghiệm cho thấy, trong những tình huống nhất định, chẳng hạn như khi bộ dữ liệu có chứa nhiều mẫu dài và dài, kỹ thuật này có thể giảm đáng kể tổng chi phí khai phá định dạng dọc của tập mục phổ biến.

6.3.2.6. Khai phá mẫu cực đại và đóng

Việc khai phá tập mục phổ biến có thể tạo ra một số lượng lớn các tập mục phổ biến, đặc biệt là khi ngưỡng min-sup thiết lập thấp hoặc khi tồn tại mẫu dài trong bộ dữ liệu. Việc dùng tập mục

phổ biến đóng, tức tập không có tập trên mà có cùng min-sup, có thể giảm đáng kể số lượng các mẫu tạo ra trong khai phá tập mục phổ biến, mà vẫn giữ các thông tin đầy đủ về tập các tập mục phổ biến. Đó là, từ tập các tập mục phổ biến đóng, có thể dễ dàng lấy được tập các tập mục phổ biến và trợ giúp của chúng. Vì vậy trong thực tế, người ta muốn khai phá tập các tập mục phổ biến đóng, hơn là tập tất cả các tập mục phổ biến.

Liên quan đến khai phá tập mục phổ biến đóng, có tiếp cận đơn giản (i) khai phá tập các tập mục phổ biến; rồi (ii) loại bỏ tất cả các tập mục phổ biến con có cùng số trợ giúp. Tuy nhiên, việc này gây tốn kém.

Một phương pháp luận được khuyến cáo là tìm kiếm trực tiếp các tập mục phổ biến, trong quá trình khai phá. Nó đòi hỏi phải xén không gian tìm kiếm ngay khi có thể định tên các trường hợp của các tập mục đóng trong khai phá. Chiến lược xén bao gồm:

Định nghĩa: Hợp nhất mục phát biểu nếu mọi giao dịch có chứa một tập mục phổ biến X cũng chứa tập mục phổ biến Y , nhưng không chứa bất kỳ tập trên của Y , thì $X \cup Y$ tạo thành tập mục phổ biến đóng, và không cần tìm kiếm bất kỳ tập mục phổ biến có chứa X nhưng không chứa Y .

Thí dụ cơ sở dữ liệu điều kiện chiếu trong thí dụ trên, đối với tập mục tiền tố $\{I_5:2\}$ là $\{\{I_2, I_1\}, \{I_2, I_1, I_3\}\}$, cho phép người ta thấy rằng các giao dịch của chúng chứa tập mục $\{I_2, I_1\}$, nhưng không chứa tập trên của $\{I_2, I_1\}$. Tập mục $\{I_2, I_1\}$ có thể hợp với $\{I_5\}$ để tạo nên tập mục đóng $\{I_5, I_2, I_1:2\}$, và không cần khai phá các tập mục đóng chứa $\{I_5\}$, nhưng không $\{I_2, I_1\}$.

Định nghĩa: Xén tập mục con phát biểu nếu một tập mục phổ biến X là một tập con thực sự của tập mục phổ biến Y đóng đã phát hiện, và $số\ trợ\ giúp(X) = số\ trợ\ giúp(Y)$, thì X và tất cả các con X trong cây liệt kê tập không thể là các tập mục đóng phổ biến, do đó có thể xén đi.

Nếu cơ sở dữ liệu giao dịch chỉ có hai giao dịch $\{(a_1, \dots, a_{100}), (a_1, \dots, a_{50})\}$ và min-sup = 2. Chiếu trên mục a_1 thu được tập mục phổ

biến $\{a_1, \dots, a_{50}:2\}$ dựa vào tối ưu *hợp nhất mục*. Do trợ giúp($\{a_2\}$) = trợ giúp($\{a_1, \dots, a_{50}\}$) = 2, và $\{a_2\}$ là tập con thực sự của $\{a_1, \dots, a_{50}\}$, người ta không cần kiểm tra a_2 và cơ sở dữ liệu chiều của nó. Tương tự, việc xén được thực hiện với $a_3, \dots, a_{50}$. Do vậy, khai phá các tập mục đóng phổ biến trong bộ dữ liệu này kết thúc sau khi khai phá cơ sở dữ liệu chiều của a_1 .

Định nghĩa: Bỏ qua mục phát biểu trong khai phá sâu của các tập mục đóng, ở mỗi cấp, sẽ có tập mục tiên tố X liên kết với bảng tiêu đề và cơ sở dữ liệu chiều. Nếu mục phổ biến địa phương p có cùng trợ giúp trong một số bảng tiêu đề ở mức khác nhau, có thể xén p từ các bảng tiêu đề, ở mức cao hơn.

Chẳng hạn cơ sở dữ liệu giao dịch trước đó chỉ có hai giao dịch $\{(a_1, \dots, a_{100}), (a_1, \dots, a_{50})\}$ và $\text{min-sup} = 2$. Do a_2 trong cơ sở dữ liệu chiều của a_1 có cùng trợ giúp như a_2 trong bảng tiêu đề chung, có thể xén a_2 khỏi bảng tiêu đề chung. Tương tự, việc xén được thực hiện với $a_3, \dots, a_{50}$. Do vậy, khai phá các tập mục đóng phổ biến trong bộ dữ liệu này kết thúc sau khi khai phá cơ sở dữ liệu chiều của a_1 .

Bên cạnh việc xén không gian tìm kiếm trong quá trình khai phá tập mục phổ biến đóng, việc tối ưu hóa quan trọng là thực hiện kiểm tra hiệu quả mỗi tập mục phổ biến mới được rút ra, xem nó là đóng hay không. Sở dĩ như vậy là do quá trình khai phá không thể đảm bảo rằng tất cả các tập mục phổ biến được tạo ra là đóng.

Khi rút ra một tập mục phổ biến mới, cần thiết thực hiện hai loại kiểm tra tính đóng: (i) *kiểm tra tập trên*, tức kiểm tra tập mục phổ biến mới này là tập trên của các tập mục đóng đã phát hiện có cùng trợ giúp không; (ii) *kiểm tra tập con*, tức tập mục mới phát hiện là tập con của tập mục đã phát hiện với cùng trợ giúp không.

Nếu chấp nhận phương pháp xén *hợp nhất mục*, trong khung chia để trị, thì việc kiểm tra tập trên được thực sự đã có, và không cần thực hiện tường minh việc kiểm tra tập trên. Lí do là tập mục

phổ biến $X \cup Y$ được phát hiện sau tập mục X , và có cùng trợ giúp như X , nó cần có trong cơ sở dữ liệu chiếu của X , và cần được tạo trong lúc hợp nhất mục.

Để giúp kiểm tra tập con, người ta xây dựng *cây mẫu*, để quản lí tập các tập mục đóng. Về cấu trúc, cây mẫu tương tự như cây FP, ngoại trừ tất cả các tập mục đóng phát hiện được lưu tương minh trong nhánh cây tương ứng. Để kiểm tra tập con hiệu quả, người ta sử dụng tính chất sau.

Tính chất đóng: Nếu tập mục phổ biến hiện tại S_c nằm trong tập mục đóng đã phát hiện S_a , thì (i) S_c và S_a có cùng trợ giúp; (ii) độ dài S_c nhỏ hơn độ dài S_a ; (iii) tất cả các mục trong S_c nằm trong S_a .

Dựa vào tính chất này, *cấu trúc chỉ mục băm hai mức* được xây dựng để truy cập nhanh cây mẫu: (i) mức đầu dùng để định danh mục cuối cùng trong S_c như khóa băm, do định danh này cần nằm trong nhánh S_c ; (ii) mức hai sử dụng trợ giúp của S_c như khóa băm, do S_c và S_a có cùng trợ giúp. Việc này đẩy nhanh đáng kể quá trình kiểm tra tập con.

6.3.3. Các phương pháp đánh giá mẫu

Hầu hết các thuật toán khai phá luật kết hợp sử dụng khung *trợ giúp, tin cậy*. Mặc dù ngưỡng trợ giúp và tin cậy tối thiểu giúp loại bỏ việc khai phá các luật không phù hợp, nhiều luật tạo ra vẫn thích hợp với người dùng. Việc này xảy ra khi *khai phá với ngưỡng trợ giúp thấp* hoặc *khai phá các mẫu dài*. Đây là một trở ngại lớn cho ứng dụng khai phá luật kết hợp.

6.3.3.1. Luật mạnh không hẳn tốt

Một luật được đánh giá hoặc theo chủ quan, hoặc theo khách quan. Cuối cùng, chỉ có người dùng có thể kiểm chứng một luật, và đánh giá này là chủ quan. Tuy nhiên, độ đo khách quan, dựa trên thống kê trên dữ liệu, có thể được sử dụng như một bước cuối.

Để biết các luật kết hợp mạnh có ý nghĩa ra sao: xét luật kết hợp mạnh không thích hợp. Giả sử cần phân tích các giao dịch tại *Hàng điện tử*, liên quan đến việc mua trò chơi máy tính và video. Lấy g tham chiếu đến các giao dịch có trò chơi máy tính, và v tham chiếu giao dịch có mua video. Trong 10,000 giao dịch phân tích, dữ liệu cho thấy 6.000 giao dịch khách hàng có trò chơi, 7,500 có video, và 4.000 có cả hai. Giả sử chương trình khai phá dữ liệu để phát hiện luật kết hợp, với $\text{min-sup} = 30\%$ và tin cậy tối thiểu là 60% . Luật kết hợp tìm thấy là $\text{mua } (X, \text{trò chơi máy tính}) \Rightarrow \text{mua } (X, \text{video})$ ($\text{trợ giúp} = 40\%$, $\text{tin cậy} = 66\%$). Luật này là luật mạnh, do trợ giúp của nó là $4,000 / 10,000 = 40\%$ và tin cậy là $4,000 / 6,000 = 66\%$, thỏa mãn các ngưỡng. Tuy nhiên có sai lầm, do xác suất mua video là 75% , lớn hơn 66% . Trong thực tế, trò chơi máy tính và video tương quan ngược, vì việc mua thứ này sẽ giảm lượng mua thứ kia. Do thiếu hiểu biết đầy đủ, người ta dễ sa vào luật không thông minh như vừa khai phá.

Thí dụ này cũng cho thấy sự tin cậy của một luật $A \Rightarrow B$ có thể bị sai. Nó không đo *sức mạnh thực sự*, tức thiếu sức mạnh, về mối tương quan và sự *ngầm định* giữa A và B . Do đó, người ta cần giải pháp khác khung trợ giúp , tin cậy , trong khai phá các mối quan hệ dữ liệu.

6.3.3.2. Từ phân tích kết hợp đến phân tích tương quan

Độ đo trợ giúp và tin cậy chưa đủ để lọc các luật kết hợp có ý nghĩa. Để khắc phục điểm yếu này, có thể dùng độ đo tương quan, bổ sung cho khung trợ giúp , tin cậy .

Định nghĩa: Luật tương quan là luật có dạng $A \Rightarrow B$ (trợ giúp , tin cậy , *tương quan*).

Luật tương quan được đo không chỉ với trợ giúp và tin cậy , mà còn tương quan giữa các tập mục A và B . Có nhiều tương quan khác nhau. Sau đây là các độ đo tương quan, tốt đối với khai phá các bộ dữ liệu lớn.

*Độ cao*¹ là độ đo tương quan. Khi xuất hiện của tập mục A *độc lập* với B, $P(A \cup B) = P(A) \cdot P(B)$; ngược lại chúng là *phụ thuộc* và có tương quan. Xác định *độ cao* $(A, B) = P(A \cup B) / P(A) \cdot P(B)$. Nếu giá trị nhỏ hơn 1, thì việc xảy ra A là *tương quan ngược* với việc xảy ra B. Nếu giá trị lớn hơn 1, thì A và b có *tương quan dương*, với nghĩa việc xảy ra cái này sẽ ngầm việc xảy ra của cái kia. Giá trị là 1 khi A, B *độc lập*, không có tương quan.

Công thức độ cao tương đương với $P(A | B) / P(B)$, hay $c(A \Rightarrow B) / s(B)$, cũng được tham chiếu như *độ cao* của sự kết hợp, hay tương quan, của luật $A \Rightarrow B$. Nói khác đi, nó đánh giá mức độ mà việc xảy ra của thứ này sẽ *nâng cao* việc xảy ra của thứ kia.

Thí dụ 6.8 về phân tích tương quan với độ đo *độ cao*. Để giúp lọc ra kết hợp mạnh không thích hợp, cần nghiên cứu cách hai tập mục A, B tương quan nhau. Lấy g₁ là các giao dịch không có trò chơi máy tính, v₁ là không có video. Các giao dịch được tóm tắt trong *bảng ngẫu nhiên*.

Bảng 6.7. Bảng ngẫu nhiên tóm tắt về các giao dịch, với việc mua trò chơi và video

	<i>g</i>	<i>g₁</i>	<i>Tổng</i>
<i>v</i>	4.000	3.500	7.500
<i>v₁</i>	2.000	500	2.500
Tổng	6.000	4.000	10.000

Theo dữ liệu trong bảng, xác suất mua trò chơi $P(\{g\}) = 0.60$, xác suất mua video $P(\{v\}) = 0.75$ và xác suất mua cả hai $P(\{g, v\}) = 0.40$. Theo công thức tính độ cao của luật *mua* $(X, \text{trò chơi máy tính}) \Rightarrow \text{mua}(X, \text{video})$ (*trợ giúp* = 40%, *tin cậy* = 66%) là $P(\{g, v\}) / (P(\{g\}) \cdot P(\{v\})) = 0.40 / (0.60 \cdot 0.75) = 0.89 < 1$. Do giá trị nhỏ hơn 1, có tương quan ngược giữa việc xảy ra {g} và {v}. Từ số cho biết khả năng khách hàng mua cả hai, trong khi mẫu số cho biết khả năng hai

¹ Lift: thang nâng, độ cao nâng lên.

việc mua là hoàn toàn độc lập. Tương quan ngược như vậy không thể xác định theo khung trợ giúp, tin cậy.

Tương quan thứ hai được đề cập là độ đo χ^2 . Để tính giá trị này, người ta lấy bình phương của sai khác giữa giá trị quan sát và giá trị kì vọng đối với một khuôn, tức cặp A và B, trong bảng ngẫu nhiên. Lượng này là tổng tất cả các khuôn của bảng ngẫu nhiên.

Bảng 6.8. Bảng ngẫu nhiên với các giá trị kì vọng

	<i>g</i>	<i>g₁</i>	<i>Tổng</i>
<i>v</i>	4,000 (4,500)	3,500 (3,000)	7,500
<i>v₁</i>	2,000 (1,500)	500 (1,000)	2,500
Tổng	6,000	4,000	10,000

Về phân tích tương quan nhờ χ^2 : để tính toán tương quan theo χ^2 , đối với dữ liệu định danh, cần giá trị quan sát và giá trị kì vọng, thể hiện trong dấu ngoặc trong bảng, đối với mỗi khuôn của bảng ngẫu nhiên. Theo bảng, tính được χ^2 theo:

$$\chi^2 = \sum_{\text{Khuôn}} \frac{(\text{quan sát} - \text{kì vọng})^2}{\text{kì vọng}} = \frac{(4000 - 4500)^2}{4500} + \frac{(3500 - 3000)^2}{3000} + \frac{(2000 - 1500)^2}{1500} + \frac{(500 - 1000)^2}{1000} = 555.6$$

Do giá trị χ^2 lớn hơn 1, và giá trị quan sát của khuôn (*g, v*) = 4000, nhỏ hơn giá trị kì vọng 4.500, *g* và *v* có tương quan ngược.

6.3.3.3. So sánh các độ đo đánh giá mẫu

Thay vì sử dụng khung trợ giúp, tin cậy, tiếp cận sử dụng độ cao và χ^2 cho phép lộ ra nhiều mối quan hệ nội tại. Vấn đề tiếp theo là hiệu quả của các độ đo.

Nhiều nghiên cứu đã đề cập các biện pháp đánh giá mẫu, ngay cả trước khi bắt đầu nghiên cứu chuyên sâu về phương pháp mở rộng cho khai phá mẫu phổ biến. Gần đây, một số biện pháp đánh giá mẫu khác đã thu hút được sự quan tâm. Sau đây là bốn

độ đo (i) tin cậy tất cả; (ii) tin cậy cực đại; (iii) Kulczynski; (iv) cosin, và so sánh chúng với tiếp cận sử dụng *độ cao* và χ^2

Cho hai tập mục, A và B, độ đo *tin cậy tất cả* của A và B được xác định: *tin cậy tất cả*(A, B) = $\frac{s(A \cup B)}{\max\{s(A), s(B)\}} = \min\{P(A|B), P(B|A)\}$;

trong đó $\max\{s(A), s(B)\}$ là trợ giúp cực đại của tập mục A và B. Do vậy tin cậy tất cả (A, B) cũng là tin cậy tối thiểu của hai luật kết hợp liên quan đến A, B, tức $A \Rightarrow B, B \Rightarrow A$.

Độ đo tin cậy cực đại của A và B được xác định: tin cậy cực đại (A, B) = $\max\{P(A|B), P(B|A)\}$. Độ đo này là tin cậy cực đại của hai luật kết hợp $A \Rightarrow B, B \Rightarrow A$.

Độ đo Kulczynski¹ của A và B được xác định: $Kulc(A, B) = (P(A|B) + P(B|A)) / 2$. Nó cho biết trung bình của độ đo tin cậy.

Độ đo cosin của A và B được xác định: $cosin(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) * P(B)}} = \frac{s(A \cup B)}{\sqrt{s(A) * s(B)}} = \sqrt{P(A|B) * P(B|A)}$.

Độ đo cosin được coi như độ đo *độ cao hài hòa*: do *độ cao* (A, B) = $P(A \cup B) / P(A) * P(B)$, nên hai công thức tương tự, chỉ khác về dấu căn trong công thức cosin. Tuy nhiên, đó là sai khác quan trọng, do khi lấy căn, giá trị cosin chỉ ảnh hưởng từ trợ giúp của A, B, $A \cup B$, chứ không từ số tổng cộng các giao dịch.

Mỗi độ đo trong bốn độ đo trên có tính chất sau: giá trị của nó chỉ chịu ảnh hưởng từ trợ giúp của A, B, và $A \cup B$, hay chính xác hơn, từ xác suất có điều kiện $P(A|B)$, $P(B|A)$, mà không chịu ảnh hưởng từ số các giao dịch. Tính chất chung khác là giá trị của độ đo thuộc (0, 1); khi giá trị càng cao, quan hệ giữa A và B càng chặt.

Để so sánh bốn độ đo trên và *độ cao* và χ^2 so sánh sáu độ đo đánh giá mẫu trên bộ dữ liệu điển hình. So sánh các biện pháp đánh giá sáu mẫu trên bộ dữ liệu điển hình. Mối quan hệ giữa việc mua hai mục hàng sữa, kí hiệu m, và cà phê, kí hiệu c, có thể thấy trong bảng ngẫu nhiên, cho biết số các giao dịch mua mc, tức mua cả sữa và cà phê.

¹ Do nhà toán học Ba Lan, S. Kulczynski đề xuất năm 1927.

Bảng 6.9. Bảng ngẫu nhiên về hai mục

	Sữa (m)	Không mua sữa (m_1)	Tổng
Cà phê (c)	mc	m_1c	C
Không mua cà phê (c_1)	mc_1	m_1c_1	C_1
Tổng	M	M_1	

Bảng sau cho thấy tập các bộ dữ liệu giao dịch ứng với bảng ngẫu nhiên, và các giá trị kết hợp đối với mỗi độ đo.

Bảng 6.10. So sánh sáu độ đo cùng sử dụng bảng ngẫu nhiên, với nhiều bộ dữ liệu

Bộ dữ liệu	mc	m_1c	mc_1	m_1c_1	χ^2	Độ cao	Tín cậy tất cả	Tín cậy cục đại	Kulc.	cosin
D_1	10000	1000	1000	100000	90557	9.26	0.91	0.91	0.91	0.91
D_2	10000	1000	1000	100	0	1	0.91	0.91	0.91	0.91
D_3	100	1000	1000	100000	670	8.44	0.09	0.09	0.09	0.09
D_4	1000	1000	1000	100000	24740	25.75	0.5	0.5	0.5	0.5
D_5	1000	100	10000	100000	8173	9.18	0.09	0.01	0.5	0.29
D_6	1000	10	100000	100000	965	1.97	0.01	0.99	0.5	0.10

Xét bốn dòng đầu tiên, ứng với $D_1.. D_4$, người ta thấy m và c có tương quan dương trong D_1, D_2 , tương quan ngược trong D_3 và trung lập trong D_4 . Đối với D_1 và D_2 , m và c tương quan dương do m_1 , tức 10.000, lớn hơn m_1c , tức 1.000, và mc_1 , tức 1.000. Về trực quan, người ta mua sữa, tức $M = 10.000 + 1.000 = 11.000$, thì có nhiều khả năng họ mua cà phê, tức $mc / m = 10 / 11 = 91\%$, và ngược lại.

Kết quả của bốn độ đo mới cho thấy m và c có tương quan dương mạnh trong cả hai bộ dữ liệu, do có giá trị 0.91. Tuy nhiên

nâng cao và χ^2 cho giá trị đo rất khác đối với D_1 và D_2 , do tính nhạy cảm với m_{ic1} . Thực tế trong nhiều kịch bản, m_{ic1} thường lớn và không ổn định. Ví dụ, trong cơ sở dữ liệu mua hàng, tổng số lượng giao dịch có thể dao động theo ngày và vượt quá số giao dịch có chứa bất kỳ tập mục phổ biến nào. Do đó, độ đo không nên bị ảnh hưởng từ các giao dịch không chứa các tập mục quan tâm; nếu không, nó sẽ tạo ra kết quả không ổn định, như minh họa trong D_1 và D_2 .

Tương tự, với D_3 , bốn độ đo cho thấy một cách chính xác rằng m và c tương quan ngược mạnh, vì tỉ lệ m / c bằng tỉ lệ mc / m , tức $100 / 1.100 = 9.1\%$. Tuy nhiên, nâng cao và χ^2 phát biểu ngược lại và sai: các giá trị của chúng đối với D_2 là giữa D_1 và D_3 .

Đối với bộ dữ liệu D_4 , cả độ đo nâng cao và χ^2 cho thấy tương quan dương cao giữa m và c , trong khi các độ đo khác cho thấy trung lập, do tỉ lệ mc / m_{ic} bằng tỉ lệ mc / mc_{ic} , là 1. Điều này có nghĩa nếu khách hàng mua cà phê, hoặc sữa, xác suất mà họ cũng sẽ mua sữa, hoặc cà phê, đúng là 50%.

Liên quan đến lí do nâng cao và χ^2 là kém khi phân biệt mối quan hệ kết hợp mẫu trong các bộ dữ liệu, người ta xem xét các *giao dịch rỗng*¹.

Định nghĩa: Giao dịch rỗng là giao dịch không chứa bất kì tập mục đang xem xét.

Trong thí dụ trên, m_{ic1} thể hiện các giao dịch rỗng. Độ đo nâng cao và χ^2 khó phân biệt các mối quan hệ kết hợp mẫu vì chúng chịu ảnh hưởng mạnh từ m_{ic1} . Thông thường, số các giao dịch rỗng có thể lớn hơn số mua hàng cá nhân, do nhiều người có thể không mua sữa lẫn cà phê. Mặt khác, bốn độ đo khác có chỉ số tốt về kết hợp mẫu, do chúng được xác định không chịu ảnh hưởng của m_{ic1} , tức không bị ảnh hưởng từ các giao dịch rỗng.

¹ Null - transaction: giao dịch rỗng.

Như vậy cần có một độ đo mà giá trị là độc lập với các giao dịch rỗng. Độ đo là *bất biến rỗng* nếu giá trị của nó không bị ảnh hưởng từ các giao dịch rỗng. Nó là tính chất quan trọng để đo các mẫu kết hợp trong cơ sở dữ liệu giao dịch lớn. Trong số sáu độ đo, chỉ nâng cao và χ^2 là không bất biến rỗng.

Liên quan đến đánh giá bốn độ đo, ngoài độ đo nâng cao và χ^2 , người ta xét tỉ lệ *bất cân đối* IR^1 , cho biết sự mất cân đối giữa hai tập mục, A và B. Nó được xác định:

$$IR(A, B) = \frac{|s(A) - s(B)|}{s(A) + s(B) - s(A \cup B)}, \text{ trong đó tử số là giá trị tuyệt đối}$$

của sai khác giữa trợ giúp của tập mục A và B, và mẫu số là số giao dịch có chứa A hoặc B. Nếu hai phía liên quan như nhau, giá trị này là 0. Ngược lại, giá trị càng lớn cho biết sự mất cân đối càng lớn. Tỉ lệ này độc lập với số giao dịch rỗng và độc lập với số các giao dịch.

Về so sánh các độ đo bất biến rỗng trong đánh giá mẫu: dù bốn độ đo là bất biến rỗng, chúng vẫn cho kết quả khác nhau đối với vài bộ dữ liệu tinh tế. Xét bộ dữ liệu D_5 và D_6 , trong đó hai sự kiện m và c có xác suất điều kiện không cân bằng, tức tỉ lệ mc / c lớn hơn 0, 9. Điều này có nghĩa rằng việc xuất hiện c đi cùng với xuất hiện của m . Tỉ lệ mc / m nhỏ hơn 0,1, cho biết m ngầm định c ít khả năng xuất hiện. Các độ đo *tin cậy tất cả* và *cosin* xem cả hai trường hợp như kết hợp ngược; và độ đo Kulc xem cả hai là trung lập. Độ đo *tin cậy cực đại* cho biết sự kết hợp dương mạnh đối với các trường hợp này. Tức các kết quả rất đa dạng.

Cần biết độ đo nào phản ánh mối quan hệ thực sự giữa việc mua sữa và cà phê, người ta xét *độ lệch cân bằng* của dữ liệu. Từ một quan điểm, chỉ $mc / (mc + mc_1) = 1000 / (1000 + 10,000) = 9,09\%$ các giao dịch liên quan đến sữa chứa cà phê trong D_5 và tỉ lệ này là $1000 / (1000 + 100,000) = 0,99\%$ trong D_6 , cả hai chỉ ra kết hợp ngược. Mặt khác, 90, 9% các giao dịch trong D_5 , tức $mc / (mc + mc_1) = 1000 / (1000 + 100)$ và 9% trong D_6 , tức $1000 / (1000 + 10)$ có

¹ IR Imbalanced ratio: tỉ lệ bất cân đối.

cà phê và sữa, tức chỉ ra kết hợp dương giữa sữa và cà phê. Những kết luận là rất khác nhau.

Đối với độ lệch cân bằng, để xử lí khi nó trung lập, như Kulc thực hiện, và trong thời gian đó chỉ ra độ lệch của nó theo IR. Đối với D_4 , người ta có $IR(m, c) = 0$, một trường hợp hoàn toàn cân bằng; đối với D_5 , $IR(m, c) = 0, 89$, một trường hợp khá cân bằng; trong khi đó đối với D_6 , $IR(m, c) = 0, 99$, một trường hợp rất sai lệch. Do đó, hai độ đo Kulc và IR làm việc cùng nhau sẽ thể hiện bức tranh rõ ràng cho tất cả ba bộ dữ liệu, $D_4.. D_6$.

Tóm lại, việc sử dụng chỉ độ đo trợ giúp và tin cậy, để khai phá kết hợp có thể tạo nên số lớn các luật, nhiều trong số đó có thể không thích hợp đối với người dùng. Thay vì vậy, người ta tăng cường khung trợ giúp, tin cậy, bằng độ đo mẫu, giúp tập trung khai phá với các mối quan hệ mẫu mạnh. Các độ đo bổ sung cho phép giảm đáng kể số các luật tạo ra và cho phép phát hiện các luật ý nghĩa hơn.

6.3.4. Kết luận về khai phá luật kết hợp

- Việc phát hiện các mẫu phổ biến, kết hợp, mối quan hệ tương quan trong lượng lớn dữ liệu là hữu ích trong tiếp thị chọn lọc, phân tích quyết định, và quản trị kinh doanh. Ứng dụng thông thường là *phân tích giỏ hàng*, tức nghiên cứu thói quen khách hàng, nhờ tìm kiếm các tập mục thường xảy ra cùng nhau;
- *Khai phá luật kết hợp* bao gồm (i) phát hiện *tập mục phổ biến*, tức tập các mục hàng, A và B, đáp ứng ngưỡng trợ giúp, hoặc tỉ lệ phần trăm của các bộ hợp lí; (ii) tạo các luật kết hợp mạnh $A \Rightarrow B$. Các luật này cũng thỏa mãn ngưỡng tin cậy tối thiểu, tức xác suất đã xác định về xuất hiện B với điều kiện xuất hiện A. Luật kết hợp có thể được tiếp tục phân tích để phát hiện ra *luật tương quan*, cho biết tương quan thống kê giữa tập mục A và B;

- Nhiều thuật toán hiệu quả và thuật toán mở rộng đã được phát triển để *khai phá tập mục phổ biến*, từ đó rút ra luật kết hợp và luật tương quan. Các thuật toán này có thể được phân thành ba loại: (i) thuật toán như Apriori; (ii) thuật toán tăng trưởng mẫu phổ biến; (iii) thuật toán sử dụng định dạng dữ liệu dọc;
- *Thuật toán Apriori* là một thuật toán gốc, cho phép khai phá tập mục phổ biến cho luật kết hợp logic. Nó khai phá tính chất khai phá theo mức, nói rằng tất cả *tập con khác rỗng của một tập phổ biến cũng phải là phổ biến*. Tại lần lặp thứ k , $k \geq 2$, nó tạo ứng viên k tập mục phổ biến dựa trên $(k - 1)$ tập mục phổ biến và duyệt cơ sở dữ liệu một lần để tìm ra tập đầy đủ của k - tập mục phổ biến, L_k . Biến thể liên quan đến việc băm và giảm giao dịch được sử dụng để tăng hiệu quả. Các biến thể khác bao gồm (i) phân đoạn dữ liệu, tức khai phá trên mỗi đoạn rồi hợp các kết quả; (ii) lấy mẫu dữ liệu, tức khai phá trên tập con mẫu. Những biến thể này cho phép giảm số lần duyệt dữ liệu cần thiết;
- *Tăng trưởng mẫu phổ biến* là một phương pháp khai thác tập phổ biến mà không tạo các ứng viên. Nó xây dựng một cấu trúc dữ liệu rất nhỏ gọn, tức cây FP, để nén cơ sở dữ liệu giao dịch ban đầu. Thay vì sử dụng các chiến lược *tạo và thử* của các phương pháp như Apriori, nó tập trung vào phát triển mẫu, tức đoạn, phổ biến, tránh chi phí để tạo các ứng viên;
- *Khai phá tập mục phổ biến nhờ sử dụng định dạng dữ liệu dọc*, tức phương pháp Eclat, là phương pháp biến đổi bộ dữ liệu về các giao dịch theo định dạng dữ liệu ngang tên tập mục TID, thành các định dạng dữ liệu dọc theo TID. Nó khai phá bộ dữ liệu đã chuyển nhờ thực hiện giao bộ TID, áp dụng tính chất Apriori và các kỹ thuật tối ưu như *tập các sai khác*;

- Không phải tất cả các luật kết hợp mạnh đều tốt. Do đó, khung theo trợ giúp, tin cậy cần được tăng cường với độ đo đánh giá mẫu, tốt cho khai phá luật có ý nghĩa. Một độ đo là *bất biến rộng* nếu giá trị của nó không bị ảnh hưởng từ các *giao dịch rộng*, tức giao dịch không chứa bất kỳ các tập mục đang xem xét. Trong số nhiều độ đo, có *nâng cao*, χ^2 , *tin cậy tất cả*, *tin cậy cực đại*, *Kulczynski* và *consin*. Người ta thấy nên sử dụng độ đo Kulczynski, cùng với tỉ lệ mất cân đối, để thể hiện các mối quan hệ mẫu trong các tập mục.

6.4. Kết luận

Với nhan đề luật kết hợp, chương này bắt đầu đề cập nội dung về thuật toán khai phá dữ liệu. Thuật toán luật kết hợp là một trong ba lớp thuật toán (i) khai phá luật kết hợp; (ii) phân lớp; (iii) phân cụm. Do vậy đầu chương đề cập (i) tổng quan về nhiệm vụ khai phá dữ liệu; (ii) sơ bộ về thuật toán khai phá dữ liệu, theo quan điểm hệ thống.

Sau phần giới thiệu chung, liên quan đến cả ba lớp thuật toán khai phá dữ liệu, chương này tập trung trình bày khai phá luật kết hợp. Nội dung này được chi tiết với (i) các phương pháp khai phá tập mục phổ biến; (ii) các phương pháp đánh giá mẫu. Các thí dụ minh họa liên quan đến việc bán, mua hàng điện tử.

Phân lớp và phân cụm

7.1. Về phân lớp và phân cụm

Việc phân lớp và phân cụm không khác nhau nhiều. Liên quan đến khai phá dữ liệu, người ta tập trung vào (i) khai phá luật kết hợp; (ii) phân lớp; (iii) phân cụm. Do tính tương đồng của phân lớp và phân cụm, chúng được trình bày trong cùng một chương, tuy nhiên, mỗi phần đều có các thuật toán khai phá dữ liệu riêng.

Định nghĩa: Phân lớp là loại phân tích thống kê, rút ra mô hình mô tả lớp dữ liệu quan trọng.

Phân lớp và phân cụm có quy trình tương tự, nhưng khác nhau về ý nghĩa. Theo góc nhìn của khai phá dữ liệu, phân lớp xác định trước số các lớp; phân cụm xác định trước số cụm. Theo góc nhìn của học máy, phân loại là kỹ thuật học có giám sát, xếp các đối tượng mới vào lớp đã huấn luyện; phân cụm là kỹ thuật học không giám sát, nhóm các đối tượng vào cụm theo đặc trưng tương tự nhau.

7.2. Phân lớp

Mô hình phân lớp, tức bộ phân lớp, dự đoán các nhãn lớp theo phạm trù, hay rời rạc, không thứ tự. Nhiều phương pháp phân lớp có trong (i) học máy; (ii) nhận dạng; (iii) thống kê. Hầu hết các thuật toán đặt trên bộ nhớ chính, do kích thước nhỏ; tuy

gần đây người ta phát triển các kĩ thuật xử lí số lượng lớn dữ liệu trên bộ nhớ ngoài.

Phân lớp có nhiều ứng dụng, kể như (i) phát hiện lỗi; (ii) tiếp thị đích; (iii) dự đoán hiệu năng; (iv) sản xuất; (v) chẩn đoán y khoa.

7.2.1. Khái niệm cơ bản

Sau một số khái niệm cơ bản, phần này trình bày tiếp cận tổng quan để phân lớp theo hai bước. (i) Bước đầu xây dựng mô hình phân lớp dựa trên dữ liệu có trước; (ii) bước sau xác định mức chính xác của mô hình và sử dụng mô hình trong phân lớp dữ liệu mới.

7.2.1.1. Khái niệm về phân lớp

Cán bộ ngân hàng cần phân tích dữ liệu để biết tình trạng hoạt động ngân hàng là an toàn hay rủi ro. Cán bộ tiếp thị hàng điện tử cần dữ liệu để phân tích lượng khách mua hàng. Trong các thí dụ này, nhiệm vụ phân tích dữ liệu là *phân lớp*, trong đó mô hình, hay *bộ phân lớp* được kiến thiết để dự đoán các *nhãn lớp*. Các nhãn này là *an toàn* hay *rủi ro*, *có* hay *không*. Các phạm trù này có thể thể hiện qua các giá trị rời rạc, mà thứ tự trong các giá trị không có ý nghĩa.

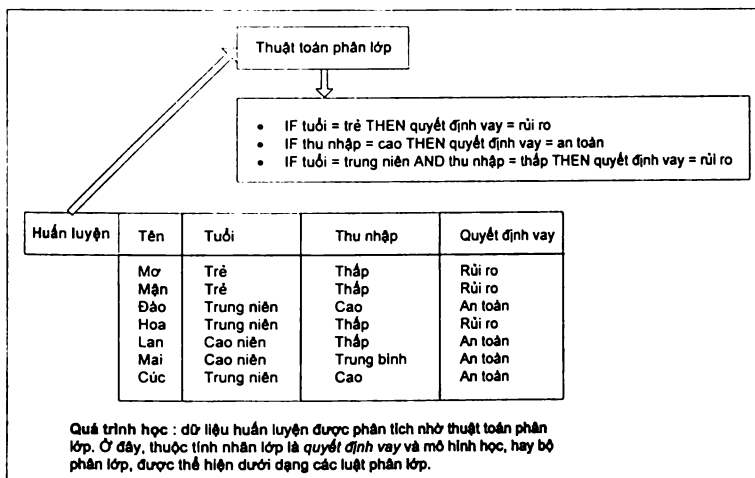
Giả sử cán bộ tiếp thị muốn dự đoán số lượng khách đến mua hàng điện tử. Nhiệm vụ phân tích dữ liệu này là *dự đoán số*, khi đó mô hình được kiến thiết dự đoán *hàm giá trị liên tục*, hay *giá trị có thứ tự*, trái với nhãn lớp. Mô hình này được gọi là *bộ dự đoán*. Phân tích hồi quy là phương pháp luận thống kê, thường dùng để dự đoán số. Hai loại chính của bài toán dự đoán là (i) phân lớp; (ii) dự đoán số. Ở đây đề cập bài toán phân lớp.

7.2.1.2. Tiếp cận tổng quát về phân lớp

Phân lớp dữ liệu theo hai giai đoạn (i) *giai đoạn học*, kiến thiết mô hình; (ii) *giai đoạn phân lớp*, lúc này mô hình dùng để dự đoán

các nhãn lớp trên dữ liệu đã cho. Quá trình trên ứng dụng vay tiền trong hình.

Trong bước thứ nhất, bộ phân lớp được xây dựng khi mô tả tập các lớp dữ liệu, hay các khái niệm. Đây là bước học, hay pha huấn luyện, trong đó thuật toán phân lớp xây dựng bộ phân lớp bằng cách phân tích, hay học, trên *tập huấn luyện*. Tập huấn luyện là các *dòng*¹ cơ sở dữ liệu và các nhãn thể hiện mối quan hệ giữa các dòng. Một dòng là $X = (x_1, \dots, x_n)$, là *vector thuộc tính n chiều*, mô tả n số đo trên dòng từ n thuộc tính cơ sở dữ liệu $A_1, \dots, A_n$. Giả sử mỗi dòng dữ liệu thuộc về một lớp xác định trước. Lớp này do thuộc tính cơ sở dữ liệu khác quyết định, gọi là *thuộc tính nhãn lớp*. Thuộc tính nhãn lớp có giá trị rời rạc và không có thứ tự. Chính nhờ các *phạm trù*, hay định danh, mà mỗi giá trị dùng như một phạm trù, hay lớp. Các dòng dùng trong pha này gọi là các dòng *huấn luyện*, được lấy mẫu ngẫu nhiên khi phân tích cơ sở dữ liệu. Trong ngữ cảnh phân lớp, các dòng dữ liệu được tham chiếu như *mẫu, thể hiện, điểm dữ liệu*, hay *đối tượng*.

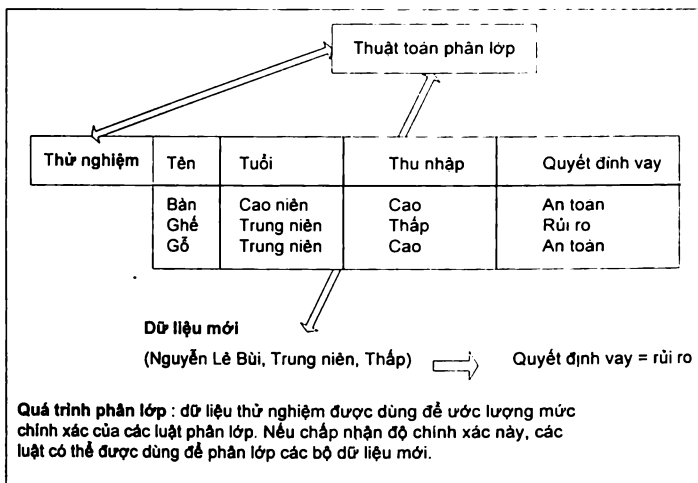


Hình 7.1. Quá trình học

¹ Tuple: bộ dữ liệu theo mô hình quan hệ, hoặc gọi là dòng dữ liệu; ở đây sử dụng thuật ngữ *dòng*; tránh lẫn với *bộ dữ liệu* (data set).

Do nhãn lớp của mỗi huấn luyện *đã cho*, pha này được biết như *học có giám sát*, tức việc học máy của bộ phân lớp là giám sát được, theo nghĩa đã biết dòng dữ liệu nào thuộc về lớp nào. *Học không giám sát*, hay *phân cụm*, chưa biết nhãn lớp, và số hay tập các lớp cần phát hiện ra là chưa biết trước.

Pha đầu của quá trình phân lớp được xem như quá trình ánh xạ $y = f(X)$, dự đoán nhãn lớp y của dòng X đã biết. Người ta muốn học cách ánh xạ, tức hàm, để tách các lớp dữ liệu. Điển hình là ánh xạ có dạng luật phân lớp, cây quyết định, công thức toán học. Trong hình thí dụ, có các luật phân lớp. Các luật này dùng để tạo phạm trù cho dữ liệu tương lai, cũng như để nghiên cứu sâu hơn về dữ liệu. Chúng cũng thể hiện dữ liệu một cách cô đọng.



Hình 7.2. Quá trình phân lớp

Trong pha thứ hai, mô hình này dùng để phân lớp. Đầu tiên người ta đánh giá độ chính xác dự đoán của bộ phân lớp. Nếu sử dụng tập dữ liệu huấn luyện để đo mức chính xác thì không khách quan, do dữ liệu huấn luyện có xu hướng phù hợp với bộ phân lớp. Người ta sử dụng thuật ngữ *quá khít*¹, tức khi học nó kết hợp vài điểm bất thường của tập huấn luyện, mà không thể hiện trong

¹ Overfit: quá khít.

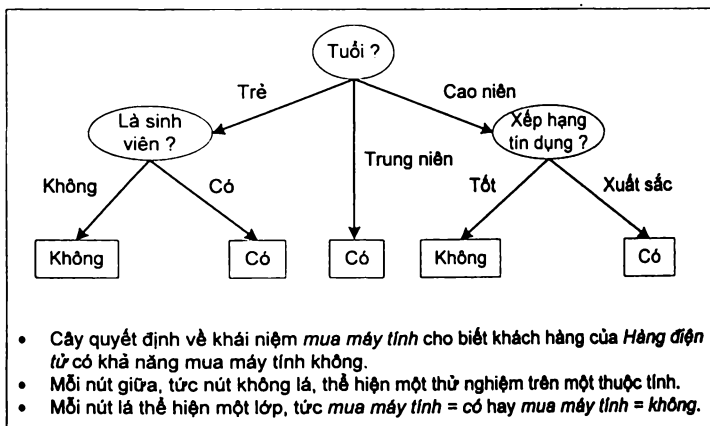
toàn thể dữ liệu. Do vậy, cần sử dụng *tập thử nghiệm*, tạo từ các *dòng dữ liệu thử* và các *nhân lớp* liên quan.

Độ chính xác¹ của bộ phân lớp trên tập thử nghiệm là tỉ lệ dòng dữ liệu phân lớp đúng. Nhân lớp liên quan của mỗi dòng thử nghiệm được so sánh với dự đoán của về lớp của bộ phân lớp. Nếu chấp nhận độ chính xác này, người ta dùng bộ phân lớp cho dữ liệu tương lai, tức trường hợp chưa biết nhân lớp.

7.2.2. Cây quyết định

Định nghĩa: Quy nạp cây quyết định là việc học từ các dòng dữ liệu huấn luyện; cây quyết định có cấu trúc như cây, trong đó nút trung gian cho biết phép thử trên thuộc tính, mỗi nhánh thể hiện một kết quả thử nghiệm, mỗi nút lá ứng với nhân lớp.

Nút trên cùng của cây quyết định là nút gốc. Hình sau là cây quyết định điển hình, với khái niệm *mua máy tính*; cây dự đoán khách hàng mua máy tính hay không. Nút trung gian kí hiệu bằng hình chữ nhật, nút lá hình ô van. Vài thuật toán cây quyết định sinh ra chỉ cây nhị phân, nhưng có thuật toán sinh ra cây tổng quát.



Hình 7.3. Cây quyết định

¹ Accuracy: độ chính xác (ngoài ra còn thuật ngữ chính xác chủ quan (precision); chính xác khách quan, hay độ phủ (recall)).

Về cách dùng cây quyết định để phân lớp, người ta cho dòng X , mà chưa biết nhãn lớp liên quan, thử các giá trị thuộc tính theo cây quyết định. Như vậy sẽ hình thành đường đi đến nút lá; tại đó có lớp chứa dòng dữ liệu này. Cây quyết định có thể chuyển thành các luật.

Cây quyết định là thông dụng do vài lí do. (i) Kiến thiết bộ phân lớp theo cây quyết định không đòi hỏi tri thức lĩnh vực nào hay thiết lập tham số, nên phù hợp với khám phá tri thức. (ii) Các cây quyết định có thể chứa dữ liệu nhiều chiều. (iii) Biểu diễn tri thức theo dạng cây là trực quan và tương tự theo cách con người nghĩ. (iv) Các pha học và phân lớp theo quy nạp cây quyết định là đơn giản và nhanh. (v) Nói chung cây quyết định có độ chính xác chấp nhận được. Tuy nhiên thành công của quá trình phân lớp còn tùy vào dữ liệu sẵn có. Các thuật toán quy nạp cây quyết định dùng để phân lớp trong nhiều lĩnh vực ứng dụng, như *y tế*, sản xuất, phân tích tài chính, thiên văn, và sinh học phân tử. Các cây quyết định là cơ sở của vài hệ thống quy nạp dùng tri thức luật.

Tiếp theo là thuật toán học cây quyết định. Khi kiến thiết cây, *độ đo lựa chọn thuộc tính* dùng để chọn thuộc tính có khả năng phân biệt tốt. Khi đã có cây quyết định, nhiều nhánh phản ánh dữ liệu lẻ hay dữ liệu ồn trong dữ liệu huấn luyện. Việc *xén cây* cho phép định tên và loại bỏ các nhánh cây đó, nhằm tăng cường độ chính xác của phân lớp trên bộ dữ liệu chưa biết. Các khía cạnh mở rộng quy nạp cây quyết định đối với cơ sở dữ liệu lớn có thể dùng tiếp cận khai phá đồ họa.

7.2.2.1. Quy nạp cây quyết định

Cuối những năm 70, đầu 80 của thế kỉ trước, J. Ross Quinlan đã phát triển thuật toán ID3¹ cho cây quyết định. Kết quả này kế thừa các hệ thống học khái niệm trước đó của E. B. Hunt... Quinlan phát triển tiếp C4.5, cho phép các thuật toán học có giám

¹ ID3: Interactive Dichotomiser: bộ phân lớp lặp.

sát về sau đem so sánh như khung thử nghiệm. Năm 1984, nhóm các nhà thống kê mô tả thể hệ các cây quyết định nhị phân, trong cuốn sách *cây hồi quy và phân lớp CART*¹. Hai hướng phát triển độc lập, nhưng đều dựa vào dòng dữ liệu huấn luyện.

Thuật toán ID3, C4.5 và CART chấp nhận tiếp cận tham lam, tức không quay lui; cây quyết định được xây dựng theo cách chia để trị, từ trên xuống. Hầu hết các thuật toán đối với cây quyết định cũng theo tiếp cận trên xuống, bắt đầu với tập các dòng huấn luyện và các nhãn lớp liên quan. Tập huấn luyện được tách đệ quy thành các tập con. Hình 7.3. là thuật toán cây quyết định cơ bản.

Thuật toán được gọi với ba tham số D , *danh sách thuộc tính* và *phương pháp chọn thuộc tính*. D là phân² dữ liệu; ban đầu gồm các dòng dữ liệu huấn luyện và nhãn lớp liên quan. *Danh sách thuộc tính* là danh sách các thuộc tính mô tả dòng dữ liệu. *Phương pháp chọn thuộc tính* là thủ tục may rủi, để chọn thuộc tính có khả năng phân biệt tốt nhất các dòng dữ liệu đã biết vào các lớp. Thủ tục này dùng độ đo lựa chọn thuộc tính như (i) lợi ích về thông tin; (ii) chỉ số Gini. Phép đo lựa chọn này xác định chắc chắn cây được chia thành hai. Nếu không cây sẽ có nhiều nhánh.

Cây bắt đầu với nút gốc đơn, N , thể hiện các dòng dữ liệu huấn luyện D , tức pha đầu.

Thuật toán: tạo cây quyết định, cho phép sinh cây quyết định từ các dòng dữ liệu huấn luyện D .

Đầu vào:

- Phần dữ liệu, D , là tập các dòng huấn luyện và nhãn lớp của chúng;
- *Danh sách thuộc tính*, là tập các thuộc tính ứng viên;

¹ CART: classification and regression tree: phân lớp và cây hồi quy.

² Partition: phần, đoạn (trong phân đoạn); ở đây sử dụng lẫn nhau.

- *Phương pháp lựa chọn thuộc tính*, là thủ tục xác định tiêu chí tách tốt nhất đối với các dòng dữ liệu thành các lớp riêng. Tiêu chí này gồm *thuộc tính tách* và *điểm tách*, hay *tập con tách*.

Đầu ra: một cây quyết định.

Phương pháp:

Tạo nút N;

If các dòng trong D là cùng lớp, C, then

Return N là nút lá, với nhãn là lớp C;

If danh sách thuộc tính là rỗng then

Return N là nút lá được đánh nhãn với lớp đa số trong D; / *bầu đa số

Gọi phương pháp chọn thuộc tính (D, danh sách thuộc tính) để tìm tiêu chí tách tốt nhất;

Đánh nhãn nút N với tiêu chí tách;

If thuộc tính tách có giá trị rời rạc and được phép tách nhiều cách then / * không hạn chế cây nhị phân

Danh sách thuộc tính: = danh sách thuộc tính - thuộc tính tách; / * loại bỏ thuộc tính tách

For each kết quả j của tiêu chí tách / * phân đoạn các dòng và tạo tập con đối với mỗi phần

D_j: = tập các dòng dữ liệu trong D thoả mãn kết quả j; / * một phân đoạn

If D_j là rỗng then gán một danh sách nhãn với lớp đa số trong D cho nút N;

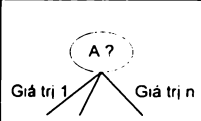
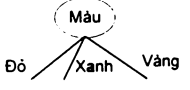
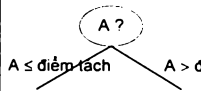
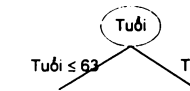
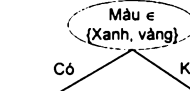
Else gán nút trả về từ tạo cây quyết định (D_j, danh sách thuộc tính) cho nút N;

Endfor

Return N;

Thuật toán cơ bản đối với quy nạp câu quyết định từ các dòng huấn luyện.

Nút N được gán nhãn theo tiêu chí tách, dùng như thử nghiệm tại nút này. Một nhánh phát sinh từ N đối với mỗi kết quả của tiêu chí tách. Các dòng dữ liệu trong D được phân đoạn tương ứng. Có ba kịch bản có thể, như minh họa trong hình. Cho A là thuộc tính tách; A có v giá trị phân biệt $\{a_1, \dots, a_v\}$ dựa trên dữ liệu huấn luyện.

A là thuộc tính phân tách		
Kịch bản phân đoạn	Thí dụ minh họa	
		Nếu A có giá trị rời rạc, thì mỗi nhánh được phát triển ứng với mỗi giá trị A .
		Nếu A có giá trị liên tục, thì hai nhánh tiến triển ứng với $A \leq$ điểm tách và $A >$ điểm tách.
		Nếu A có giá trị rời rạc và cần theo cây nhị phân, thì thử nghiệm có dạng $A \in S_A$, trong đó S_A là tập con tách, đối với A .

Hình 7.4. Ba khả năng phân đoạn, dựa trên tiêu chí tách khác nhau

1. A có giá trị rời rạc. kết quả của thử nghiệm tại nút N ứng trực tiếp đến giá trị đã biết của A . Tạo được một nhánh đối với mỗi giá trị đã biết, a_i , của A và được gán nhãn với giá trị đó. Phần D_i là tập con của các dòng được gán nhãn lớp trong D với giá trị a_i của A . Do tất cả các dòng trong phần đã cho có cùng giá trị đối với A , A không cần xét trong bất kì phân đoạn sắp tới của các dòng. Do vậy, nó bị loại ra khỏi danh sách thuộc tính.

2. *A có giá trị liên tục.* Thử nghiệm tại nút A có hai kết quả, ứng với điều kiện $A \leq \text{điểm tách}$ và $A > \text{điểm tách}$. Điểm tách là *điểm tách* do *phương pháp chọn thuộc tính* trả về. Thực tế, điểm tách A thường được lấy là điểm giữa của hai giá trị kế nhau của A; do vậy có thể không có mặt trong A từ dữ liệu huấn luyện, hai nhánh sinh ra từ N và có nhãn theo kết quả trước đó. Các dòng dữ liệu được phân đoạn như D_1 giữ tập con của các dòng có nhãn lớp trong D đối với $A \leq \text{điểm tách}$, và D_2 giữ phần còn lại;
3. *A có giá trị rời rạc và cần sinh cây nhị phân.* Thử nghiệm tại nút N có dạng $A \in S_A?$; trong đó S_A là tập con tách đối với A, do *phương pháp chọn thuộc tính* trả về như một phần của tiêu chí tách. Đó là tập con các giá trị đã biết của A. Nếu một dòng đã cho có giá trị a_i của A và nếu $a_i \in S_A$, thử nghiệm này tại nút n là thỏa mãn. Hai nhánh sinh ra từ N. Theo lệ thường, nhánh trái từ N được gán nhãn *có*¹, để D_1 ứng với tập con của các dòng có nhãn lớp trong D mà thỏa mãn phép thử. Nhánh phải từ N có nhãn *không*, để D_2 ứng với tập con của các dòng có nhãn lớp trong D mà không thỏa mãn phép thử.

Độ phức tạp về tính toán của thuật toán này với tập huấn luyện D là $O(n * |D| * \log(|D|))$, trong đó n là số các thuộc tính trong D, và $|D|$ là số các dòng huấn luyện trong D. Điều này có nghĩa giá tính toán tăng theo phát triển của cây, nhiều nhất đến $n * |D| * \log(|D|)$, với $|D|$ dòng dữ liệu.

Thế hệ gia tăng của quy nạp cây quyết định theo hướng: cho tập huấn luyện mới, việc tái cấu trúc cây quyết định thu được từ việc học trên dữ liệu trước, thay vì học lại cây mới.

7.2.2.2. Độ đo lựa chọn thuộc tính

Định nghĩa: Độ đo chọn thuộc tính là kinh nghiệm, may rủi, để chọn tiêu chí tách, sao cho tách tốt nhất phân dữ liệu đã cho, gồm các dòng dữ liệu huấn luyện đã có nhãn lớp thành các lớp riêng.

¹ Yes: có.

Nếu muốn tách D thành các phần nhỏ hơn, theo kết quả của tiêu chí tách, lí tưởng là mỗi phần cần sạch, tức tất cả các dòng dữ liệu rơi vào phần đã cho sẽ thuộc cùng một lớp. Về khái niệm, tiêu chí tách tốt nhất là tiêu chí mà kết quả sát nhất theo kịch bản. Các độ đo chọn thuộc tính được xem là các *luật tách*, do chúng xác định cách các dòng tại một nút được tách.

Độ đo chọn thuộc tính đảm bảo phân hạng đối với mỗi thuộc tính có trong các dòng dữ liệu. Thuộc tính này có điểm tốt nhất sẽ được chọn làm *thuộc tính tách* đối với tập các dòng dữ liệu đã cho. Nếu thuộc tính tách có giá trị liên tục, hay nếu người ta muốn hạn chế ở cây nhị phân, thì hoặc *điểm tách*, hoặc *tập con tách* được xác định như một phần của tiêu chí tách. Nút cây tạo ra cho phần D được đánh nhãn theo tiêu chí tách; các nhánh lớn lên theo kết quả của tiêu chí này, và người ta phân đoạn các dòng. Sau đây là ba độ đo chọn thuộc tính (i) lợi ích thông tin; (ii) tỉ suất lợi; (iii) chỉ số Gini.

Các kí pháp sẽ thống nhất là: phân dữ liệu D , là tập huấn luyện của các dòng đã có nhãn lớp. Giả sử thuộc tính nhãn lớp này có m giá trị phân biệt, xác định m lớp khác nhau C_i , $i = 1..m$. Cho $C_i \cap D$ là tập các dòng của lớp C_i trong D . Kí hiệu $|D|$ và $|C_i \cap D|$ là số các dòng.

Lợi ích thông tin: thuật toán ID3 sử dụng *lợi ích thông tin*¹ như độ đo chọn thuộc tính. Độ đo dựa trên lí thuyết thông tin, với C . Shannon, nghiên cứu giá trị hay nội dung thông tin của thông điệp. Cho nút N thể hiện, hay giữ, các dòng dữ liệu của đoạn D . Thuộc tính có nhiều lợi ích về thông tin sẽ được chọn như thuộc tính tách đối với nút N . Thuộc tính này tối thiểu thông tin cần thiết, để phân lớp các dòng dữ liệu trong các phần kết quả, và phản ánh mức ngẫu nhiên ít nhất, hay mức bẩn, tại các phần. Tiếp cận như vậy sẽ tối thiểu hóa số các phép thử cần thiết để phân lớp dòng dữ liệu đã biết và bảo đảm phát hiện được cây đơn giản.

¹ Information gain: lợi ích về thông tin.

Thông tin kì vọng cần để phân lớp một dòng dữ liệu trong D là $Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$; trong đó p_i là xác suất khác không, để dòng bất kì trong D thuộc về lớp C_i , và được đánh giá là $|C_i \cap D| / |D|$. Hàm $\log_2$ có mặt do thông tin mã hóa theo bit. $Info(D)$ đúng là lượng trung bình về thông tin cần để định danh nhãn lớp của một dòng trong D . Lưu ý thông tin này chỉ dựa trên tỉ lệ các dòng của mỗi lớp. $Info(D)$ được biết như *entropy* của D .

Giả sử phân đoạn các dòng dữ liệu trong D trên thuộc tính A nào đó có v giá trị phân biệt $\{a_1, \dots, a_v\}$, như được quan sát từ tập huấn luyện. Nếu A có giá trị rời rạc, các giá trị này trực tiếp tương ứng với v kết quả của thử nghiệm trên A . Thuộc tính A có thể được sử dụng để tách D thành v tập con, hay các phần $\{D_1, \dots, D_v\}$, trong đó D_j chứa các dòng trong D có kết quả a_j của A . Các phần này sẽ tương ứng với các nhánh sinh ra từ nút N . Lí tưởng là việc phân đoạn này phân lớp chính xác các dòng dữ liệu. Tức là người ta muốn các phần là sạch. Tuy nhiên các phần này không sạch, tức một đoạn gồm cả tập các dòng từ các lớp khác, chứ không của một lớp.

Lượng thông tin người ta cần, sau phân đoạn, để có phân lớp chính xác có số đo $Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} * Info(D_j)$. Lượng $|D_j| / |D|$ có vai trò trọng số cho phần thứ j ; $Info_A(D)$ là thông tin kì vọng cần để phân lớp một dòng từ D , dựa trên việc A phân đoạn. Kì vọng càng thấp, thì độ sạch càng cao.

Định nghĩa: Lợi ích thông tin được xác định như hiệu giữa yêu cầu thông tin gốc, tức chỉ dựa vào tỉ lệ của các lớp, với yêu cầu mới, tức sau khi phân đoạn.

Vậy $lợi(A) = Info(D) - Info_A(D)$. Nói khác đi, lợi (A) cho biết ích lợi ra sao nếu rẽ nhánh trên A . Đó chính là mức giảm kì vọng về yêu cầu thông tin, có được do biết giá trị của A . Thuộc tính A với độ đo lợi ích lớn nhất, lợi (A), được chọn làm thuật toán tách tại nút N . Nó có nghĩa rằng muốn phân đoạn trên thuộc tính A mà

phân lớp tốt nhất, sao cho lượng thông tin cần để phân lớp các dòng là tối thiểu, tức tối thiểu hóa $\text{Info}_A(D)$.

Về quy nạp cây quyết định với lợi ích thông tin: xét bảng thể hiện tập huấn luyện D , của các dòng dữ liệu có nhãn lớp, theo cơ sở dữ liệu *Hàng điện tử*. Ở đây mỗi thuộc tính có giá trị rời rạc; các giá trị liên tục cần được tổng quát hóa.

Bảng 7.1. Các dòng dữ liệu huấn luyện có nhãn lớp, từ cơ sở dữ liệu hàng điện tử

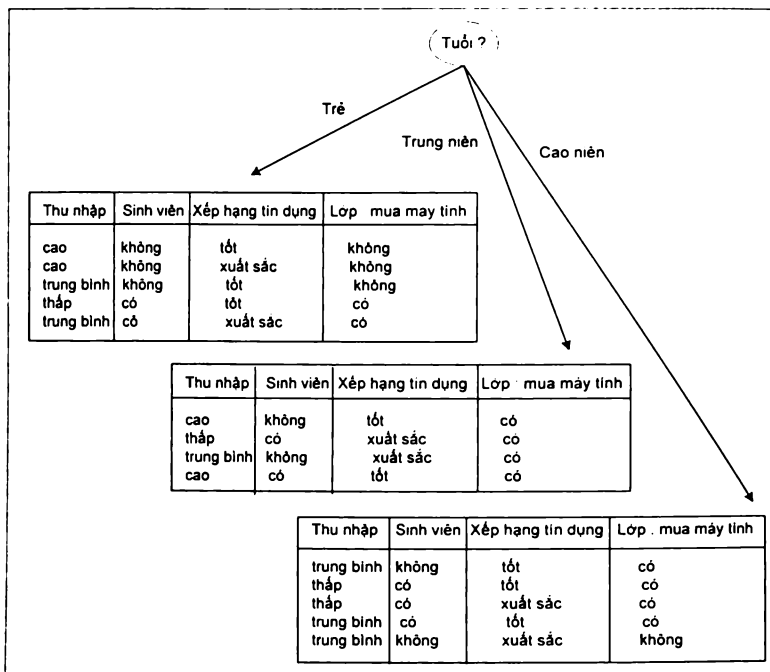
	Tuổi	Thu nhập	Sinh viên	Xếp hạng tin dụng	Lớp : mua máy tính
1	trẻ	cao	không	tốt	không
2	trẻ	cao	không	xuất sắc	không
3	trung niên	cao	không	tốt	có
4	cao niên	trung bình	không	tốt	có
5	cao niên	thấp	có	tốt	có
6	cao niên	thấp	có	xuất sắc	có
7	trung niên	thấp	có	xuất sắc	có
8	trẻ	trung bình	không	tốt	không
9	trẻ	thấp	có	tốt	có
10	cao niên	trung bình	có	tốt	có
11	trẻ	trung bình	có	xuất sắc	có
12	trung niên	trung bình	không	xuất sắc	có
13	trung niên	cao	có	tốt	có
14	cao niên	trung bình	không	xuất sắc	không

Thuộc tính nhãn lớp *mua máy tính* có hai giá trị phân biệt, lấy là {có, không}. Do vậy có hai lớp phân biệt, $m = 2$. Lấy lớp C_1 ứng với có, C_2 với không. Có 9 dòng lớp có và năm dòng lớp không. Nút gốc N được tạo cho các dòng trong D .

Để phát hiện tiêu chí tách trên các dòng dữ liệu, cần tính lợi ích thông tin đối với mỗi thuộc tính. Trước hết $\text{Info}(D) = -(9/14) \cdot \log_2(9/14) - (5/14) \cdot \log_2(5/14) = 0.940$ bit. Tiếp theo tính yêu cầu thông tin kì vọng được mỗi thuộc tính. Bắt đầu với thuộc tính *tuổi*. Thuộc tính này chia phạm trù *trẻ*, với hai dòng có và ba dòng không. Với phạm trù *trung niên*, thấy bốn có và không có không. Với phạm trù *già*, thấy ba có và hai không. Do vậy $\text{Info}_{\text{tuổi}}(D) = (5/14) \cdot (- (2/5) \cdot \log_2(2/5) - (3/5) \cdot \log_2(3/5)) + (4/14) \cdot (- (4/4) \cdot \log_2(4/4))$

+ $(5/14) * (- (3/5) * \log_2 (3/5) - (2/5) * \log_2 (2/5)) = 0.694$ bit. Do vậy lợi (tuổi) = Info (D) - Info_{tuổi} (D) - $0.940 - 0.694 = 0.246$ bit.

Tương tự, tính được lợi (thu nhập) = 0.029 bit; lợi (sinh viên) = 0.151 bit. Do lợi (tuổi) lớn nhất trong số các thuộc tính, nó được chọn như thuộc tính tách tại nút N. Nút N lấy nhãn là *tuổi*, và các nhánh sinh ra đối với mỗi giá trị thuộc tính. Các dòng chia đoạn theo *tuổi* = *trung niên* sẽ thuộc cùng lớp, như trong hình. Do tất cả chúng đều thuộc lớp *có*, lá của cây được tạo tại cuối nhánh này, và có nhãn là *có*. Toàn bộ cây quyết định đã thấy trong hình.



Hình 7.5. Các dòng dữ liệu theo phân đoạn, bắt đầu từ thuộc tính *tuổi*

Có vấn đề là đối với thuộc tính có giá trị liên tục, không tính được lợi ích thông tin như đối với thuộc tính có giá trị rời rạc. Giả sử có thuộc tính A có giá trị liên tục, chẳng hạn như độ tuổi là giá

trị số. Khi ấy cần xác định *điểm tách* đối với thuộc tính A, là ngưỡng trên A.

Trước tiên, sắp xếp các giá trị của A theo thứ tự tăng. Diễn hình là điểm giữa mỗi cặp giá trị kế nhau được xem như ứng viên của điểm tách. Chẳng hạn điểm giữa của giá trị a_i và a_{i+1} là $(a_i + a_{i+1})/2$. Khi các giá trị của A được xếp, xác định cách tách A tốt nhất đòi hỏi chỉ một lần duyệt các giá trị. Đối với mỗi điểm tách, người ta tính $\text{Info}_A(D)$. D_1 là tập các dòng dữ liệu trong D thỏa mãn $A \leq$ điểm tách, và D_2 là các dòng mà $A >$ điểm tách.

Tỉ lệ lợi ích: Độ đo lợi ích thông tin thiên về các phép thử với nhiều kết quả. Tức nó có xu hướng chọn các thuộc tính có số lớn các giá trị. Chẳng hạn xét thuộc tính có định danh duy nhất là *#sản phẩm*; tách trên nó sẽ tạo nên nhiều phân đoạn, nhiều như số các giá trị, mà mỗi đoạn chỉ có một dòng. Do mỗi đoạn là sạch, thông tin cần thiết để phân lớp bộ dữ liệu D dựa trên phân đoạn này sẽ là $\text{Info}_{\text{sản phẩm}}(D) = 0$. Do vậy thông tin được lợi nhờ phép tách theo thuộc tính này là lớn nhất. Thế mà phân đoạn này là vô dụng để phân lớp.

Thuật toán C4.5 sử dụng mở rộng của lợi ích thông tin bằng *tỉ lệ lợi ích*¹, khắc phục thiên lệch trên. Nó dùng để chuẩn hóa lợi ích thông tin, nhờ giá trị về thông tin tách, $\text{SplitInfo}_A(D) = -\sum_{j=1}^v \frac{|D_j|}{|D|}$

$\cdot \log_2 \frac{|D_j|}{|D|}$. Giá trị này thể hiện thông tin tiềm năng tạo ra nhờ tách bộ dữ liệu huấn luyện D thành v phần, ứng với v kết quả của phép thử trên thuộc tính A. Lưu ý đối với mỗi kết quả, số các dòng có kết quả đó ứng với số tổng cộng các dòng trong D. Nó khác với lợi ích thông tin: lợi ích thông tin đo thông tin ứng với phân lớp thu được, dựa trên cùng một cách phân đoạn. Tỉ lệ lợi ích là $\text{GainRatio}(A) = \text{lợi}(A) / \text{SplitInfo}_A(D)$. Thuộc tính có tỉ lệ lợi ích cao nhất sẽ được chọn để tách. Tuy nhiên khi thông tin tách dẫn đến 0, tỉ lệ

¹ Gain ratio: tỉ lệ lợi ích.

này không ổn định; vậy có hạn chế để tránh sự kiện này, theo đó lợi ích thông tin của thử nghiệm được chọn cần lớn, ít nhất cũng lớn hơn lợi ích trung bình so với tất cả thử nghiệm được xét.

Về tính tỉ lệ lợi ích đối với thuộc tính thu nhập: phép thử trên thu nhập tách dữ liệu Bảng 8.1 thành ba phần, tên là *thấp*, *vừa*, *cao*, có bốn, sáu và bốn dòng. Người ta tính $\text{SplitInfo}_{\text{thu nhập}}(D) = - (4 / 14) * \log_2 (4 / 14) - (6 / 14) * \log_2 (6 / 14) - (4 / 14) * \log_2 (4 / 14) = 1.557$. Trong thí dụ trên, lợi (*thu nhập*) = 0.029, nên $\text{GainRatio}(\text{thu nhập}) = 0.029 / 1.557 = 0.019$.

Chỉ số Gini: CART sử dụng chỉ số Gini. Chỉ số Gini dùng để đo độ bẩn của phần dữ liệu D , hay của tập các dòng dữ liệu huấn luyện $\text{Gini}(D) = 1 - \sum_{i=1}^m p_i^2$; trong đó p_i là xác suất để dòng trong D thuộc về lớp C_i , và được đánh giá bằng $|C_i \cap D| / |D|$.

Định nghĩa: Chỉ số Gini là chỉ số biểu thị độ bất bình đẳng trong phân phối thu nhập. Nó có giá trị từ 0 đến 1 và bằng tỉ số giữa phần diện tích nằm giữa đường cong Lorenz¹ và đường bình đẳng tuyệt đối với phần diện tích nằm dưới đường bình đẳng tuyệt đối.

Chỉ số Gini để ý tách nhị phân đối với mỗi thuộc tính. Trước tiên xét trường hợp A có giá trị rời rạc với v giá trị phân biệt $\{a_1, \dots, a_v\}$ xảy ra trong D . Để xác định tách nhị phân tốt nhất trên A , người ta kiểm tra tất cả các tập con có khả năng được tạo từ các giá trị chưa biết của A . Mỗi tập con S_A được xét như phép thử nhị phân được thuộc tính A ở dạng $A \in S_A$? Cho một dòng, phép thử này là thỏa mãn nếu giá trị của A đối với dòng này có trong số các giá trị liệt kê trong S_A . Nếu A có v giá trị khả năng, có thể có 2^v tập con. Chẳng hạn với thuộc tính *thu nhập* với ba giá trị *thấp*, *vừa*, *cao*; có thể có tập con $\{\text{thấp}, \text{vừa}, \text{cao}\}$, $\{\text{thấp}, \text{vừa}\}$, $\{\text{thấp}, \text{cao}\}$, $\{\text{vừa}, \text{cao}\}$, $\{\text{thấp}\}$, $\{\text{vừa}\}$, $\{\text{cao}\}$, và $\{\}$. Loại bỏ $\{\text{thấp}, \text{vừa}, \text{cao}\}$ và $\{\}$ do chúng không thể hiện việc tách. Do vậy còn $2^v - 2$ khả năng tạo nên hai phần dữ liệu D , dựa trên việc tách nhị phân trên A .

¹ Đường cong thể hiện mức bất bình đẳng trong phân phối, có từ 1905.

Khi xét việc tách nhị phân, người ta tính tổng có trọng số về độ bẩn của mỗi phần. Chẳng hạn nếu tách nhị phân trên A , từ D được D_1 và D_2 , thì chỉ số Gini của D là $Gini_A(D) = (|D_1| / |D|) * Gini(D_1) + (|D_2| / |D|) * Gini(D_2)$. Đối với mỗi thuộc tính, đều xét khả năng tách nhị phân. Đối với thuộc tính giá trị rời rạc, tập con có chỉ số Gini thấp nhất đối với thuộc tính đó sẽ được chọn như tập con tách.

Đối với thuộc tính có giá trị liên tục, cần xét từng điểm tách có thể. Cách làm tương tự như mô tả trên đối với lợi ích thông tin, mà giá trị điểm giữa các cặp số kế nhau, đã sắp xếp, là điểm tách. Điểm này, với chỉ số Gini thấp nhất đối với thuộc tính liên tục được chọn là điểm tách của thuộc tính này. Với mỗi điểm tách của A , D_1 là tập các dòng dữ liệu thỏa mãn $A \leq$ điểm tách; D_2 là tập con các dòng thỏa mãn $A >$ điểm tách.

Việc làm sạch khi tách nhị phân trên thuộc tính A rời rạc hay liên tục là $\Delta Gini(D) = Gini(D) - Gini_A(D)$. Thuộc tính cực đại hóa trong việc làm sạch Δ , tức chỉ số Gini cực tiểu, được chọn làm thuộc tính tách. Thuộc tính này và hoặc tập con tách của nó khi thuộc tính giá trị rời rạc, hoặc điểm tách khi thuộc tính là liên tục, cùng tạo nên tiêu chí tách.

Về quy nạp cây quyết định với chỉ số Gini: Cho D là tập huấn luyện như trong bảng dữ liệu; có 9 dòng thuộc về lớp *mua máy tính* = có, và năm dòng còn lại thuộc về lớp *không*. Nút gốc N tạo ra cho các dòng của D . Tính được độ bẩn của D ; $Gini(D) = 1 - (9/14)^2 - (5/14)^2 = 0.459$.

Để phát hiện tiêu chí tách cho các dòng trong D , cần tính chỉ số Gini đối với mỗi thuộc tính. Bắt đầu với thuộc tính *thu nhập*, và xét mỗi tập con có khả năng. Xét tập con {thấp, vừa}; nó cho 10 dòng trong phần D_1 , thỏa mãn *thu nhập* $\in$ {thấp, vừa}. Bốn dòng còn lại được gán cho D_2 . Giá trị chỉ số Gini được tính $Gini_{thu nhập \in \{thấp, vừa\}}(D) = (10/14) * Gini(D_1) + (4/14) * Gini(D_2) = (10/14) * (1 - (7/10)^2 - (3/10)^2) + (4/14) * (1 - (2/4)^2 - (2/4)^2) = 0.443$. Tương tự

tính được $Gini_{thu\ nh\hat{a}p \in \{th\grave{a}p, cao\}}(D) = 0.458$; $Gini_{thu\ nh\hat{a}p \in \{v\grave{u}a, cao\}}(D) = 0.450$; $Gini_{thu\ nh\hat{a}p \in \{cao\}}(D) = 0.443$; $Gini_{thu\ nh\hat{a}p \in \{th\grave{a}p\}}(D) = 0.450$; $Gini_{thu\ nh\hat{a}p \in \{v\grave{u}a\}}(D) = 0.458$. Do vậy giải pháp tốt nhất là tách theo {thấp, vừa} hay {cao} đối với thuộc tính *thu nhập*, do chúng có giá trị Gini thấp. Đánh giá *tuổi*, thấy {trẻ, già} hay {trung niên} là tách tốt nhất, với giá trị chỉ số Gini là 0.375; đánh giá *sinh viên* và *tin dụng*, đều là nhị phân, với giá trị chỉ số Gini tương ứng là 0.367 và 0.429.

Thuộc tính *tuổi* và tập con tách {trẻ, già} cho giá trị thấp nhất toàn thể, nên độ giảm bẩn, tức làm sạch, $0.459 - 0.357 = 0.102$. Tách nhị phân *tuổi* $\in \{trẻ, già\}$? làm giảm bẩn nhiều nhất của các dòng trong D , được xem như tiêu chí tách. Nút N được gán nhãn với tiêu chí này; hai nhánh sinh ra từ đó, và các dòng được phân đoạn tương ứng.

Độ đo lựa chọn thuộc tính khác: nhận xét về ba độ đo trình bày trên (i) độ đo lợi ích thông tin thiên về thuộc tính đa giá trị; (ii) độ đo tỉ lệ lợi ích, khắc phục độ đo trước, có xu hướng về tách không cân bằng, mà có phần nhỏ hơn phần khác; (iii) chỉ số Gini thiên về các thuộc tính đa giá trị, phức tạp khi có nhiều lớp; ngoài ra, nó có xu hướng thuận lợi cho phép thử mà kết quả là các phần kích thước như nhau và sạch. Tuy vậy chỉ số Gini tốt trong thực tế.

Thuật toán cây quyết định CHAID dùng độ đo chọn thuộc tính dựa trên phép thử thống kê χ^2 đối với tính độc lập.

Các độ đo chọn thuộc tính dựa trên nguyên lí độ dài mô tả tối thiểu MDL¹ ít thiên lệch về các thuộc tính đa giá trị. Độ đo dựa trên MDL dùng kĩ thuật mã hóa để xác định cây quyết định tốt nhất, theo cách cần ít bit mã hóa nhất để (i) mã hóa cây; (ii) mã hóa phần còn lại, không thuộc cây.

Có độ đo xét việc *tách nhiều biến*, tức phân đoạn các dòng dựa vào *tổ hợp* của các thuộc tính, thay vì một thuộc tính. Việc này tạo nên *cấu trúc thuộc tính*, hay *đặc trưng*, xây dựng từ các thuộc tính đã biết.

¹ MDL: Minimum Description length: độ dài mô tả tối thiểu.

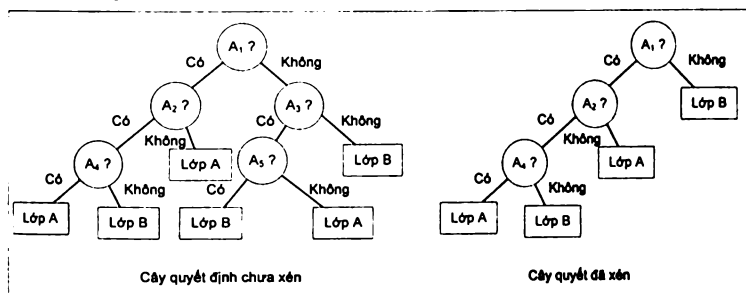
Việc chọn độ đo nào tùy vào (i) độ phức tạp về thời gian tăng theo hàm mũ đối với độ cao của cây; (ii) hình dạng phức tạp của cây, tức cây nhị phân hay nhiều nhánh.

7.2.2.3. Xén cây

Các nhánh cây sinh ra do *quá khít*, tức thể hiện các dữ liệu mẫu mà không có trong thực tế, sẽ bị cắt bỏ. Việc này thu gọn cây, tăng tốc độ phân lớp đối với dữ liệu thực. Có hai tiếp cận thông dụng (i) xén trước; (ii) xén sau¹.

Theo tiếp cận *xén trước*, cây ngừng phát triển, tức quyết định không tách tiếp các dòng dữ liệu huấn luyện tại một nút. Như vậy nút này trở thành nút lá; nút lá này thể hiện lớp hay xảy ra nhất trong số các dòng dữ liệu con, hay phân phối xác suất của các dòng này. Khi kiến thiết cây, độ đo như ý nghĩa thống kê, lợi ích thông tin, chỉ số Gini... được dùng để đánh giá việc tách tốt. Nếu việc phân đoạn các dòng rơi vào dưới ngưỡng thì người ta ngừng việc tách; xác định ngưỡng cũng là vấn đề.

Tiếp cận thông dụng hơn là *xén sau*, loại bỏ cây con từ cây chính. Như vậy, nút xóa cây con sẽ trở thành nút lá; nhận nhãn lớp thường xảy ra.



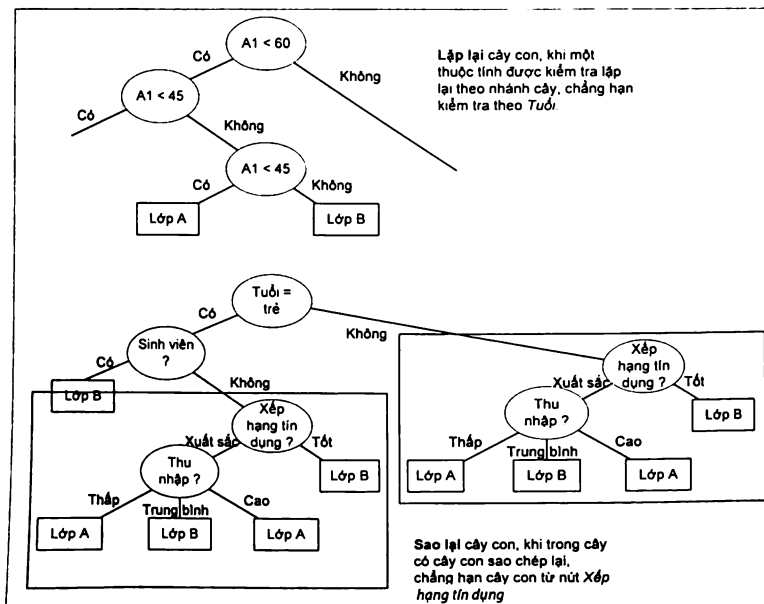
Hình 7.6. Cây quyết định chưa xén, và cây đã xén

Một số nguyên lý để xén cây (i) thuật toán CART sử dụng *độ phức tạp về giá*; (ii) thuật toán C4.5. sử dụng *xén bị quan*. Nguyên lý

¹ Prepruning, postpruning: xén trước, xén sau.

này dựa vào đánh giá tỉ lệ sai sót để ra quyết định. Thay vì xén cây dựa vào tỉ lệ sai sót, có thể xén theo số các bit cần để mã hóa chúng. Cây đã xén tốt nhất là cây tối thiểu hóa số bit mã hóa. Phương pháp này chấp nhận nguyên lí MDL; ý tưởng cơ bản là dùng giải pháp đơn giản nhất. Nó không cần tập các dòng dữ liệu độc lập, như tiếp cận xén theo giá.

Dù cây xén có khuynh hướng cô đọng so với nguyên gốc, chúng vẫn lớn và phức tạp. Các cây quyết định có thể xuất hiện *lặp* hay *bản sao*¹, như trong hình, gây khó diễn giải. Lặp xảy ra khi thuộc tính được thử lần nữa theo nhánh đã biết, chẳng hạn *tuổi* < 60? sau khi đã qua *tuổi* < 45?. Trường hợp sao lại xuất hiện khi lặp lại cây con đã có. Việc này làm mất tính chính xác và nhận thức của cây quyết định. Sử dụng tách nhiều biến, tức kết hợp thuộc tính, sẽ khắc phục được tình trạng bản sao.



Hình 7.7. Các cây con

¹ Repetition: lặp; replication: bản sao.

7.2.2.4. Mở rộng cây quyết định

Khi quy nạp cây quyết định, số lượng lớn dòng dữ liệu không thể tải trong bộ nhớ trong được. Dù năng lực máy tính tăng, nhưng dung lượng bộ nhớ trong không đủ đáp ứng bài toán thực, với hàng triệu dữ liệu. Người ta đề xuất (i) dùng bảng giá trị thuộc tính và nhãn; (ii) thuật toán tối ưu *thủ khởi động*¹ đối với kiến thiết cây.

Giá trị thuộc tính và nhãn: một trong các tiếp cận về mở rộng kích thước quy nạp cây quyết định là *rừng mưa*². Nó sử dụng bảng của tập các giá trị thuộc tính và nhãn lớp được gán, AVC³, đối với mỗi thuộc tính, tại mỗi nút của cây, mô tả các dòng huấn luyện tại nút đó. Bảng AVC của thuộc tính A tại nút n cho biết nhãn lớp đối với mỗi giá trị của A đối với các dòng tại N. Hình cho thấy bảng ứng với dữ liệu trong bảng đầu chương.

Tuổi			Thu nhập		
Mua máy tính			Mua máy tính		
	Có	Không		Có	Không
Trẻ	2	3	Thấp	3	1
Trung niên	4	0	Trung bình	4	2
Cao niên	3	2	Cao	2	2

Sinh viên			Xếp hạng		
Mua máy tính			Mua máy tính		
	Có	Không		Có	Không
Đúng	6	1	Tín dụng	6	2
Sai	3	4	Tốt	3	3
			Xuất sắc		

Sử dụng cấu trúc dữ liệu để gộp thông tin theo dữ liệu huấn luyện. Nó cho phép cải thiện phạm vi suy luận với cây quyết định.

Hình 7.8. Sử dụng cấu trúc dữ liệu

Tập tất cả các bảng AVC tại nút N là *nhóm AVC*. Kích thước của một bảng AVC tại N phụ thuộc chỉ vào số các giá trị khác nhau của A và số các lớp trong tập các dòng dữ liệu tại N. Về cơ bản, kích thước này phù hợp với bộ nhớ trong.

¹ Bootstrap: thủ khởi động, khởi động lại.

² RainForest: rừng mưa.

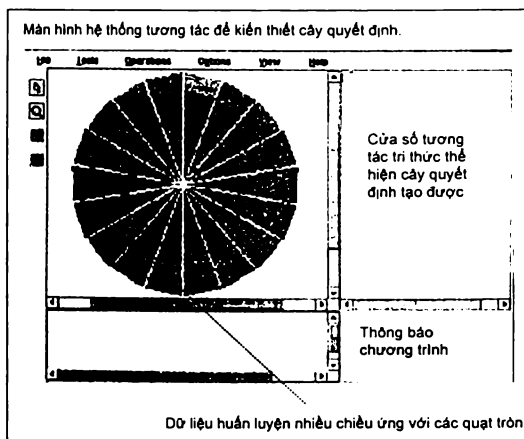
³ AVC: Attribute - Value ClassLabel: giá trị thuộc tính và nhãn lớp.

Kiến thiết cây với thuật toán tối ưu tự khởi động: tiếp cận này khác với cách mở rộng nhờ cấu trúc lại dữ liệu hay dùng cấu trúc dữ liệu đặc biệt. Thay vì tác động đến cấu trúc dữ liệu, nó dùng kĩ thuật thống kê thử khởi động.

Định nghĩa: Thử khởi động là thuật ngữ xuất xứ từ thế kỉ XIX tại Hoa Kỳ, thường dùng để chỉ một quá trình tự khởi đầu mà không cần tác động từ bên ngoài. Đối với công nghệ thông tin, nó đề cập quá trình tải các phần mềm cơ bản vào bộ nhớ của máy tính sau khi bật máy hoặc thiết lập lại, đặc biệt hệ điều hành mà thực hiện việc tải phần mềm khác.

Tiếp cận này tạo một vài mẫu nhỏ, hay tập con, của dữ liệu huấn luyện, phù hợp với kích thước bộ nhớ trong. Người ta kiến thiết một cây từ mỗi tập con này, được T1. T1 gắn với cây T cần tìm. Phương pháp kiến thiết cây BOAT¹ có thể sử dụng bất kì độ đo lựa chọn thuộc tính nào, để tách nhị phân và dựa trên tính sạch của phần dữ liệu. Việc điều chỉnh T1 để có T được thực hiện trên dữ liệu gia tăng. BOAT duyệt dữ liệu chỉ hai lần, và nhanh hơn phương pháp rừng mưa hai hay ba lần.

7.2.2.5. Khai phá dữ liệu trực quan đối với cây quyết định



Hình 7.9. Màn hình minh họa

¹ BOAT: Bootstrapped Optimistic Algorithm for Tree construction: kiến thiết cây với thuật toán tối ưu tự khởi động.

Với công cụ đồ họa, khai phá dữ liệu với đồ họa hiển thị là phù hợp với xu thế tương tác người - máy. Trong hình là thí dụ khai phá dữ liệu trực quan, với công cụ phân lớp dựa trên nhận thức PBC¹. Trong hình là các khoanh màu và danh sách phân lớp.

7.2.3. Các phương pháp phân lớp Bayes

Bộ phân lớp Bayes thuộc loại thống kê. Người ta dùng nó để dự đoán xác suất thuộc vào lớp thành viên, cũng như xác suất một dòng dữ liệu thuộc vào lớp cụ thể.

Bộ phân lớp Bayes dựa vào định lý Bayes. Trong số các thuật toán phân lớp, người ta dùng bộ *phân lớp Bayes đơn giản*². Nó có hiệu năng như bộ phân lớp dùng mạng nơron hay cây quyết định. Các bộ phân lớp Bayes cho thấy tốc độ và độ chính xác cao, khi làm việc với cơ sở dữ liệu lớn.

Bộ phân lớp Bayes đơn giản giả sử rằng tác động của giá trị thuộc tính trên một lớp đã biết là độc lập với giá trị của thuộc tính khác. Gọi giả thiết này là *độc lập có điều kiện theo lớp*. Nó cho phép đơn giản việc tính toán, nên có tên là *đơn giản*.

7.2.3.1. Định lý Bayes

Cho X là dòng dữ liệu. Theo thuật ngữ Bayes, X là sự kiện. Thông thường nó được đo bằng tập n thuộc tính. H là giả thuyết nào đó, về X thuộc vào lớp C . Đối với bài toán phân lớp, người ta muốn xác định $P(H|X)$, tức xác suất xảy ra H , khi biết sự kiện X . Nói khác đi, người ta xét xác suất X thuộc vào C , khi biết các mô tả thuộc tính của X .

$P(H|X)$ là xác suất hậu nghiệm của H với điều kiện trên X . Chẳng hạn nó thể hiện xác suất khách hàng X mua máy tính, khi biết tuổi và thu nhập của họ. Trái lại, $P(H)$ là xác suất tiên nghiệm của H . Chẳng hạn xác suất X mua máy tính, bất kể tuổi tác, thu

¹ PBC: Perception based classification: phân lớp dựa trên nhận thức.

² Naïve Bayes: thuật toán Bayes đơn giản.

nhập. Xác suất hậu nghiệm $P(H|X)$ dựa trên nhiều thông tin hơn xác suất tiên nghiệm $P(H)$; $P(H)$ độc lập với X .

Tương tự $P(X|H)$ là xác suất hậu nghiệm của X theo điều kiện trên H . Chẳng hạn xác suất X có tuổi 35..., khi biết X đã mua máy tính. $P(X)$ là xác suất tiên nghiệm của X .

Các xác suất này được ước lượng từ dữ liệu. Định lý Bayes có ích để tính xác suất hậu nghiệm $P(H|X)$ theo các xác suất kia: $P(H|X) = P(X|H) \cdot P(H) / P(X)$.

7.2.3.2. Phân lớp Bayes đơn giản

Bộ phân lớp Bayes đơn giản làm việc theo năm bước sau:

Bước 1. Lấy D là tập huấn luyện và các nhãn lớp liên quan. Thông thường, mỗi dòng dữ liệu là thể hiện của vectơ thuộc tính n chiều $X = (x_1, \dots, x_n)$, mô tả n phép đo trên dòng n thuộc tính $A_1, \dots, A_n$.

Bước 2. Giả sử có m lớp $C_1, \dots, C_m$. Cho dòng dữ liệu X , bộ phân lớp sẽ dự đoán X thuộc vào lớp C_i có xác suất hậu nghiệm lớn nhất, theo điều kiện trên X . tức là bộ phân lớp Bayes đơn giản dự đoán X thuộc vào C_i , khi $P(C_i|X) > P(C_j|X)$, $j = 1..m, j \neq i$. Lớp C_i với xác suất lớn nhất gọi là *giả thiết hậu nghiệm cực đại*.

Bước 3. Do $P(X)$ là không đổi đối với mọi lớp, chỉ $P(X|C_i) \cdot P(C_i)$ cần là lớn nhất. Nếu chưa biết các xác suất tiên nghiệm, thì thường là $P(C_1) = \dots = P(C_m)$; nên $P(X|C_i)$ cần cực đại. Ngược lại, người ta tìm cực đại của $P(X|C_i) \cdot P(C_i)$. Các xác suất tiên nghiệm về lớp có thể đánh giá theo $P(C_i) = |C_i, D| / |D|$; trong đó $|C_i, D|$ là số các dòng dữ liệu huấn luyện của lớp C_i trong D .

Bước 4. Cho bộ dữ liệu với nhiều thuộc tính; việc tính $P(X|C_i)$ sẽ tốn kém. Để giảm lượng tính toán khi ước lượng $P(X|C_i)$, người ta sử dụng giả thiết về *độc lập theo điều kiện lớp*. Giả thiết rằng các giá trị thuộc tính là độc lập có điều kiện với nhau, khi biết nhãn lớp của dòng dữ liệu này, tức không có mối quan hệ phụ thuộc

trong các thuộc tính. Trong X , thuộc tính A_k nhận giá trị x_k . Do vậy $P(X|C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) * \dots * P(x_n | C_i)$.

Người ta ước lượng dễ dàng $P(x_1|C_i)$, ..., $P(x_n|C_i)$ nhờ các dòng dữ liệu huấn luyện. Đối với mỗi thuộc tính, cần xem nó là phạm trù hay lấy giá trị liên tục. Hiện tại, việc tính $P(X|C_i)$ như sau:

4.1. Nếu A_k theo kiểu phạm trù, $P(X|C_i)$ là số các dòng dữ liệu của lớp C_i trong D mà có giá trị x_k đối với A_k , chia cho số dòng của lớp C_i trong D , tức $|C_i, D|$.

4.2. Nếu A_k có giá trị liên tục, người ta cần tính toán nhiều hơn. Thuộc tính giá trị liên tục thường giả thiết có phân phối Gauss, với trung bình μ và độ lệch chuẩn σ xác định theo:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ sao cho } P(x_k | C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}).$$

Cần tính μ_{C_i} , σ_{C_i} , tức trung bình và độ lệch chuẩn của các giá trị A_k đối với dòng dữ liệu huấn luyện của lớp C_i .

Chẳng hạn lấy $X = (35, 40.000)$, trong đó A_1 , A_2 là thuộc tính *tuổi* và *thu nhập*. Thuộc tính nhân lớp là *mua máy tính*. Nhân lớp liên quan đối với X là *có*, tức *mua máy tính = có*. Giả sử *tuổi* không rời rạc, nên nó là thuộc tính giá trị liên tục. Giả sử tập huấn luyện cho biết các khách hàng trong D mua máy tính có tuổi 35 ± 12 . Nói cách khác, đối với thuộc tính *tuổi* và lớp này, người ta có $\mu = 38$, $\sigma = 12$. Người ta có thể định lượng $x_1 = 35$ cho X , để ước lượng $P(\text{tuổi} = 35 | \text{mua máy tính} = \text{có})$.

Bước 5. Để dự đoán nhân lớp của X , $P(X|C_i) * P(C_i)$ được tính đối với mỗi C_i . Bộ phân lớp dự đoán X thuộc lớp có $P(X|C_i) * P(C_i)$ cực đại.

Liên quan đến hiệu quả của bộ phân lớp Bayes, nhiều nghiên cứu thực nghiệm tiến hành so sánh với phân lớp cây quan hệ hay mạng nơron theo cùng lĩnh vực. Về lý thuyết, các bộ phân lớp Bayes có tỉ lệ lỗi thấp so với hai bộ phân lớp kia. Tuy nhiên, do chấp nhận giả thiết không chính xác, như độc lập theo điều kiện

lớp, và thiếu thông tin xác suất, tỉ lệ lỗi của phân lớp Bayes sẽ tăng lên.

Bộ phân lớp Bayes có ích trong kiểm chứng lí thuyết cho bộ phân lớp khác, mà không sử dụng định lí Bayes một cách tường minh. Chẳng hạn theo giả thiết chắc chắn, có thể có nhiều mạng nơron và các thuật toán khớp đường cong dùng giả thuyết hậu nghiệm cực đại, như bộ phân lớp Bayes đã làm.

Về dự đoán nhân lớp với bộ phân lớp Bayes đơn giản: cần dự đoán nhân lớp của dòng dữ liệu với bộ phân lớp Bayes, trên cùng dữ liệu huấn luyện trong thí dụ trước đối với cây quyết định. Dữ liệu huấn luyện cho trong bảng đầu chương. Các dòng dữ liệu mô tả theo thuộc tính *tuổi, thu nhập, sinh viên, xếp hạng tín dụng*. Thuộc tính nhân lớp *mua máy tính* có hai giá trị {có, không}. Lấy C_1 là lớp *mua máy tính = có*; C_2 ngược lại. Dòng cần phân lớp là $X = (\text{tuổi} = \text{trẻ}, \text{thu nhập} = \text{trung bình}, \text{sinh viên} = \text{có}, \text{xếp hạng tín dụng} = \text{tốt})$. Cần cực đại $P(X|C_i) \cdot P(C_i)$ với $i = 1, 2$. Xác suất tiên nghiệm $P(C_i)$ đối với mỗi lớp được tính theo các dòng dữ liệu huấn luyện.

$$P(\text{mua máy tính} = \text{có}) = 9 / 14 = 0.643;$$

$$P(\text{mua máy tính} = \text{không}) = 5 / 14 = 0.357;$$

Để tính $P(X|C_i)$ với $i = 1, 2$, người ta tính xác suất có điều kiện:

- $P(\text{tuổi} = \text{trẻ} \mid \text{mua máy tính} = \text{có}) = 2 / 9 = 0.222;$
- $P(\text{tuổi} = \text{trẻ} \mid \text{mua máy tính} = \text{không}) = 3 / 5 = 0.600;$
- $P(\text{thu nhập} = \text{trung bình} \mid \text{mua máy tính} = \text{có}) = 4 / 9 = 0.444;$
- $P(\text{thu nhập} = \text{trung bình} \mid \text{mua máy tính} = \text{không}) = 2 / 5 = 0.400;$
- $P(\text{sinh viên} = \text{có} \mid \text{mua máy tính} = \text{có}) = 6 / 9 = 0.667;$
- $P(\text{sinh viên} = \text{có} \mid \text{mua máy tính} = \text{không}) = 1 / 5 = 0.200;$
- $P(\text{xếp hạng tín dụng} = \text{tốt} \mid \text{mua máy tính} = \text{có}) = 6 / 9 = 0.667;$
- $P(\text{xếp hạng tín dụng} = \text{tốt} \mid \text{mua máy tính} = \text{không}) = 2 / 5 = 0.400.$

Sử dụng các xác suất này, $P(X \mid \text{mua máy tính} = \text{có}) = P(\text{tuổi} = \text{trẻ} \mid \text{mua máy tính} = \text{có}) \cdot P(\text{thu nhập} = \text{trung bình} \mid \text{mua máy tính}$

= có) * P(sinh viên = có | mua máy tính = có) * P(xếp hạng tín dụng = tốt | mua máy tính = có) = $0.222 * 0.444 * 0.667 * 0.667 = 0.044$.
 Tương tự $P(X | \text{mua máy tính} = \text{không}) = 0.600 * 0.400 * 0.200 * 0.400 = 0.019$.

Để phát hiện lớp C_i có $P(X | C_i) * P(C_i)$, người ta tính $P(X | \text{mua máy tính} = \text{có}) * P(\text{mua máy tính} = \text{có}) = 0.044 * 0.643 = 0.028$; $P(X | \text{mua máy tính} = \text{không}) * P(\text{mua máy tính} = \text{không}) = 0.019 * 0.357 = 0.007$. Do vậy bộ phân lớp Bayes đơn giản dự đoán *mua máy tính = có* cho dòng dữ liệu X .

Liên quan đến các giá trị xác suất là 0, người ta ước lượng $P(X | C_i)$ như tích của $P(x_1 | C_i) * \dots * P(x_n | C_i)$, dựa trên giả thiết về độc lập điều kiện về lớp. Có thể ước lượng các xác suất này theo dữ liệu huấn luyện, như trong bước 4 của thuật toán. Người ta cần tính $P(X | C_i)$ đối với mỗi lớp C_i , $i = 1..m$, để phát hiện lớp C_i đạt giá trị cực đại. Trong trường hợp không có dòng dữ liệu huấn luyện, chẳng hạn $P(\text{sinh viên} = \text{có} | \text{mua máy tính} = \text{không}) = 0$, tức có $P(x_k | C_i) = 0$; việc này khiến $P(X | C_i) = 0$. Xác suất 0 hủy tất cả những tác động của các xác suất hậu nghiệm khác, trên C_i , tham gia vào tích. Giải pháp đối với vấn đề này: có thể giả thiết cơ sở dữ liệu huấn luyện D là lớn tới mức việc thêm một dữ liệu, để tránh 0, cũng không ảnh hưởng các đánh giá giá trị xác suất. Kỹ thuật này là *điều chỉnh Laplace*, hay *ước lượng Laplace*¹.

Về sử dụng điều chỉnh Laplace để tránh giá trị xác suất 0: giả sử đối với lớp *mua máy tính = có* trong cơ sở dữ liệu huấn luyện D chứa 1.000 dòng dữ liệu, nhưng không có dòng *thu nhập = thấp*, 990 dòng *thu nhập = trung bình*, 10 dòng *thu nhập = cao*. Các xác suất của sự kiện này là 0 , $990 / 1000 = 0.990$, $10 / 1000 = 0.010$. Với điều chỉnh Laplace, người ta thêm một dòng đối với mỗi cặp *thu nhập - giá trị*. Như vậy sẽ có xác suất $1 / 1003 = 0.001$, $991 / 1003 =$

¹ Laplacian correction, Laplace estimator: điều chỉnh Laplace, ước lượng Laplace, theo tên của nhà toán học Pháp, Pierre Laplace, 1749 - 1827.

0.988, $11 / 1003 = 0.011$. Các ước lượng xác suất đã điều chỉnh gần với thực tế, mà tránh được trường hợp 0.

7.2.4. Bộ phân lớp dựa trên luật

Định nghĩa: Bộ phân lớp dựa trên luật là bộ phân lớp mà mô hình đã học thể hiện qua tập các luật IFTHEN.

7.2.4.1. Sử dụng các luật IFTHEN để phân lớp

Các luật là cách tốt để thể hiện thông tin hay tri thức. Luật có dạng IF điều kiện THEN kết luận. Chẳng hạn có luật R1: IF *tuổi = trẻ AND sinh viên = có* THEN *mua máy tính = có*. Phần IF, hay vế trái, là điều kiện trước. Phần THEN, hay vế phải là kết quả của luật. Phần điều kiện có nhiều *phép thử thuộc tính*. Luật R1 được viết lại *tuổi = trẻ $\wedge$ sinh viên = có $\rightarrow$ mua máy tính = có*.

Nếu điều kiện trong vế trái luật là đúng, thì luật đã thỏa mãn, tức luật này *đảm bảo*¹ dòng dữ liệu. Một luật được đánh giá theo (i) đảm bảo; (ii) chính xác. Cho dòng dữ liệu X trong tập nhân lớp D, người ta xác định n_1 là số các dòng mà luật R đảm bảo; n_2 là số các dòng mà P phân lớp đúng; $|D|$ là số các dòng trong D.

Định nghĩa: Đảm bảo của luật R là tỉ số giữa số các dòng mà luật R đảm bảo và số dòng trong bộ dữ liệu được đánh nhãn lớp.

Định nghĩa: Độ chính xác của luật R là tỉ số giữa các dòng được R phân lớp đúng và số dòng được R đảm bảo.

Như vậy *đảm bảo* (R) = $n_1 / |D|$; *độ chính xác* (R) = n_2 / n_1 .

Về độ đảm bảo và độ chính xác của luật: sử dụng dữ liệu trong bảng dữ liệu đầu chương. Có các dòng dữ liệu trong cơ sở dữ liệu *hàng điện tử*. Nhiệm vụ là dự đoán khách hàng mua máy tính. Xét luật R1, đảm bảo 2 trên 14 dòng. Nó phân lớp đúng hai

¹ Cover: đảm bảo.

dòng. Do vậy *đảm bảo* ($R1$) = $2 / 14 = 14.28 \%$ và *độ chính xác* ($R1$) = $2 / 2 = 100\%$.

Liên quan đến cách hoạt động bộ phân lớp dựa trên luật, để dự đoán nhãn lớp của dòng dữ liệu X đã biết, người ta xét X = (tuổi = trẻ, thu nhập = trung bình, sinh viên = có, xếp hạng tín dụng = tốt); cần phân lớp X theo *mua máy tính*. X thỏa mãn luật $R1$, tức $R1$ được *kích hoạt*; nếu $R1$ là luật duy nhất thỏa mãn, nó *cháy*¹. Khi có nhiều luật được kích hoạt, cần có điều phối để quyết định luật nào sẽ cháy.

Khi có nhiều luật được kích hoạt, hệ thống cần có *chiến lược giải quyết xung đột*. Người ta có thể sử dụng hai chiến lược (i) xếp thứ tự theo kích thước; (ii) xếp thứ tự theo luật².

- Lược đồ xếp thứ tự theo kích thước gán ưu tiên cho luật kích hoạt có độ đo về số các thuộc tính. Luật có nhiều thuộc tính được kiểm tra sẽ được cháy;
- Lược đồ thứ tự theo luật gán ưu tiên thủ công. Thứ tự này có thể (i) xếp theo lớp; (ii) xếp theo luật. Đối với việc xếp theo lớp, người ta áp dụng cho các lớp thông dụng trước...; các luật trên cùng một lớp không cần xếp thứ tự. Đối với việc xếp theo luật, người ta gán ưu tiên theo độ đảm bảo, độ chính xác, kích thước luật; tập các luật được xem như *danh sách quyết định*.

7.2.4.2. Rút ra luật từ cây quyết định

Từ tập các dữ liệu huấn luyện, người ta xây dựng được cây quyết định. Cây quyết định là phương pháp thông dụng để phân lớp; nó dễ hiểu và có độ chính xác chấp nhận được. Với ý tưởng như cây quyết định, người ta có thể xây dựng được bộ phân lớp dựa trên luật, nhờ cây quyết định đã có.

¹ Trigger: kích hoạt; fire: cháy.

² Size ordering, rule ordering: thứ tự theo kích thước, thứ tự theo luật.

Để trích rút các luật từ một cây quyết định, người ta xem một luật được tạo theo nhánh từ gốc đến nút lá. Mỗi tiêu chí tách theo nhánh đã biết là một logic AND đối với vế trái của luật, tức phần mệnh đề IF. Nút lá ứng với dự đoán về lớp, tạo nên phần phải của luật, tức phần THEN.

Về trích rút các luật phân lớp từ cây quyết định: cây quyết định trong hình trước được chuyển thành các luật phân lớp IFTHEN nhờ theo vết các nhánh từ nút gốc đến mỗi nút lá trong cây. Các luật thu được là:

R1. IF tuổi = trẻ AND sinh viên = không THEN mua máy tính = không;

R2. IF tuổi = trẻ AND sinh viên = có THEN mua máy tính = có;

R3. IF tuổi = trung niên THEN mua máy tính = có;

R4. IF tuổi = già AND xếp hạng tín dụng = tốt THEN mua máy tính = có;

R5. IF tuổi = già AND xếp hạng tín dụng = khá THEN mua máy tính = không.

Người ta ngầm định logic OR giữa các luật được trích rút. Do các luật rút trực tiếp từ cây, chúng *loại trừ lẫn nhau* và *phủ hết* các trường hợp. *Loại trừ lẫn nhau* có nghĩa người ta không thể có xung đột về luật, do không có hai luật nào được kích hoạt bởi cùng một dòng dữ liệu. *Phủ hết* có nghĩa đã có luật đối với tổ hợp thuộc tính - giá trị; nên không cần có luật mặc định. Ở đây trật tự của các luật không quan trọng.

7.2.4.3. Quy nạp dùng luật theo thuật toán phủ tuần tự

Thuật toán: các luật IF... THEN có thể trích từ dữ liệu huấn luyện, tức không cần tạo cây quyết định, theo *thuật toán phủ tuần tự*¹. Người ta sinh ra các luật lần lượt; mỗi luật cho một lớp đã biết,

¹ Sequentially Covering Algorithm: thuật toán phủ tuần tự.

phủ tất cả những dòng dữ liệu của lớp. Thuật toán phủ tuần tự sử dụng rộng rãi trong khai phá dữ liệu các tập tách biệt của các luật phân lớp.

Có nhiều thuật toán phủ tuần tự, điển hình là AQ, CN2 và RIPPER. Chiến lược chung của chúng là: mỗi lần chỉ học một luật. Khi học một luật, các dòng dữ liệu do luật này phủ sẽ được bỏ đi. Học luật tuần tự trái với quy nạp với cây quyết định. Do đường đến nút lá trong cây quyết định ứng với một luật, người ta có thể xem quy nạp với cây quyết định như việc học đồng thời tập các luật.

Tiếp theo là thuật toán phủ tuần tự cơ bản. Tại một thời điểm, các luật được học cho một lớp. Lí tưởng là khi học một luật đối với lớp C, người ta muốn luật này phủ tất cả các dòng dữ liệu của C, và không có, hay rất ít, dòng của lớp khác. Theo cách này, các luật được học sẽ có độ chính xác cao. Các luật không cần có độ phủ cao. Vậy người ta có thể dùng nhiều luật cho một lớp, sao cho các luật khác nhau có thể phủ các dòng khác nhau trong cùng một lớp. Quá trình tiếp tục cho đến khi gặp điều kiện dừng, chẳng hạn hết các dòng huấn luyện hay chất lượng của luật thấp hơn ngưỡng người dùng quy định. Thủ tục *học - một - luật* tìm luật tốt nhất đối với lớp hiện tại, khi có bộ dữ liệu huấn luyện.

Thuật toán: phủ tuần tự, học một tập luật IFTHEN để phân lớp.

Đầu vào:

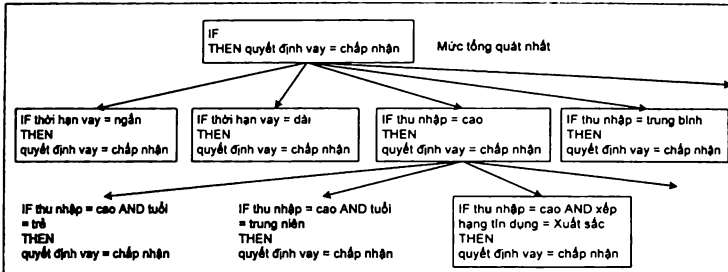
- Bộ dữ liệu D của các dòng có nhãn lớp;
- Tập các thuộc tính Att - vals và các giá trị của chúng.

Đầu ra: tập các luật IFTHEN.

Thuật toán:

```
Tập luật = {} ; // tập khởi đầu là rỗng
For each lớp C do
    Repeat
```


Luật = học - một - luật (D, att - vals, C);
 Loại bỏ khỏi D các dòng mà luật này phủ;
 Tập luật = tập luật + luật; / / bỏ sung
 luật mới
 Until điều kiện dừng;
 Endfor
 Return tập luật;



Hình 7.10. Tìm kiếm từ tổng quát đến đặc biệt trong không gian luật

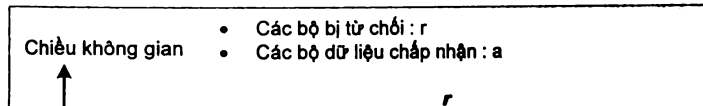
Liên quan đến cách học một luật, người ta thấy các luật phát triển theo cách từ *tổng quát đến cụ thể*, như trong hình. Với tập huấn luyện D, gồm các dữ liệu ứng dụng vay tiền. Các thuộc tính của người vay là tuổi, thu nhập, mức học vấn, nơi ở, mức tín dụng và kiểu vay. Thuật toán phân lớp là *quyết định - vay*, cho biết việc chấp thuận cho vay hay từ chối. Để học luật đối với lớp *chấp thuận*, người ta bắt đầu với luật tổng quát nhất, tức điều kiện rỗng. Luật này là IF THEN quyết định vay = chấp thuận. Rồi người ta thử thuộc tính để bổ sung vào luật này. Chẳng hạn với cặp (thuộc tính, giá trị), người ta xét thuộc tính theo thuộc tính = giá trị, thuộc tính $\leq$ giá trị, thuộc tính $>$ giá trị. Điển hình là dữ liệu huấn luyện sẽ chứa nhiều thuộc tính; mỗi thuộc tính có vài khả năng về giá trị. Việc tìm ra luật tối ưu cần đến tính toán với độ phức tạp hàm mũ. Do vậy, *học - một - luật* chấp nhận chiến lược tìm sâu. Mỗi lúc cần bổ sung thử nghiệm với thuộc tính mới cho luật hiện tại, cần chọn thuộc tính tăng cường chất lượng của luật, theo các mẫu huấn luyện. Trở lại với thí dụ trước, giả sử học - một - luật phát hiện thử

thử nghiệm thuộc tính *thu nhập* = *cao* cải thiện tốt nhất độ chính xác của luật, người ta bổ sung nó vào điều kiện, thu được *IF thu nhập cao THEN quyết định vay = chấp thuận*. Mỗi khi bổ sung thử nghiệm thuộc tính vào luật, luật kết quả sẽ phù hợp hơn các dòng luật chấp thuận. Trong lần lặp tiếp theo, xét *xếp hạng tín dụng* = *thu được IF thu nhập = cao AND xếp hạng tín dụng = tốt THEN quyết định vay = chấp thuận*.

Tìm kiếm tham lam không cho phép quay lui. Tại mỗi bước người ta bổ sung theo *kinh nghiệm*, như lựa chọn tốt.

Đo chất lượng luật: cần thiết đo chất lượng luật. Mỗi khi thử nghiệm thuộc tính, cần đảm bảo việc thêm thuộc tính sẽ cải thiện được luật.

Về chọn luật theo độ chính xác: xét hai luật như trong hình. hai luật đều dùng cho lớp *quyết định vay = chấp thuận*. Sử dụng *a* thể hiện các dòng dữ liệu của lớp *chấp thuận*, và *r* cho lớp từ chối. Luật R1 phân lớp chính xác 38 trong số 40 dòng, tức 95%. Luật chỉ phù hai dòng, chính xác 100%. Tuy R2 chính xác hơn nhưng R1 tốt hơn, do độ phù hợp lớn hơn nhiều.



Thí dụ cho thấy không tin cậy vào chỉ độ chính xác, khi đề cập chất lượng của luật. Cần có độ đo khác, cho phép tích hợp cả độ chính xác và độ phủ. Người ta đưa ra khái niệm *độ đo thông tin*, hay *lợi ích về thông tin*.

Giả sử cần học các luật cho lớp C. Có luật R1. *IF điều kiện 1 THEN lớp = C*. Khi thử thuộc tính để gia tăng điều kiện, người ta thu được R2. *IF điều kiện 2 THEN lớp = C*. Người ta hi vọng R2 tốt hơn R1.

Mức độ lợi về thông tin cũng đã được sử dụng trong lựa chọn thuộc tính trong quy nạp cây quyết định. Đó là *thông tin kì vọng* cần để phân lớp một dòng trong bộ dữ liệu D. Ở đây D là tập các dòng dữ liệu được điều kiện 2 phủ và p_i là xác suất của lớp C_i trong D. Lượng thông tin càng nhỏ thì điều kiện 2 càng tốt. Lượng thông tin hướng đến điều kiện mà phủ số lớn các dòng dữ liệu của một lớp và một ít dòng của các lớp khác.

Người ta dùng độ đo khác về mức độ lợi về thông tin là *học quy nạp vị từ FOIL*¹, là thuật toán phủ tuần tự, cho phép học các luật logic vị từ. Trong học máy, các dòng của lớp mà người ta học được các luật là các dòng *dương*; các dòng còn lại là dòng *âm*. Gọi P, N là số các dòng dương, âm do luật R phủ; và P_2 , N_2 là do luật R2 phủ. FOIL đánh giá thông tin được lợi do nói rộng điều kiện R2 là *độ lợi FOIL* $= P_2 * (\log_2 (P_2 / (P_2 + N_2)) - \log_2 (N / (N + P)))$. Có thể dùng phép thử thống kê về ý nghĩa để biết hiệu quả của một luật là không nhiều, nhưng biết tương quan rõ ràng giữa các giá trị thuộc tính và các lớp. Thử nghiệm này so sánh phân phối quan sát được trong số các lớp của các dòng do một luật phủ, với phân phối kì vọng mà sẽ thấy nếu luật này chọn dự đoán ngẫu nhiên. Người ta muốn đánh giá về bất kì sai khác quan sát được giữa hai phân phối được xem là cơ hội hay không. Người ta đề xuất *thống kê tỉ lệ có khả năng* $= 2 \sum_{i=1}^m f_i \log(\frac{f_i}{e_i})$; trong đó f_i là tần suất quan sát của mỗi lớp i trong số các dòng; e_i là điều người ta muốn về tần

¹ FOIL: First Order Inductive Learner: học quy nạp vị từ.

suất của lớp i , nếu luật này được lấy ngẫu nhiên. Thống kê có phân phối χ^2 với $m - 1$ bậc tự do. Tỷ lệ có khả năng càng cao thì càng có khả năng có *sai khác có ý nghĩa* trong số các dự đoán đúng của luật người ta dùng, so với chọn ngẫu nhiên. Tức là hiệu quả của luật người ta sử dụng không theo may mắn; tỷ lệ này giúp xác định các luật có độ phù ít ý nghĩa.

Xén luật: thủ tục học - một - luật không dùng bộ dữ liệu thử khi đánh giá luật. Việc đánh giá chất lượng luật đã mô tả được thực hiện với các dòng từ dữ liệu huấn luyện gốc. Các đánh giá này là lý tưởng do các luật có vẻ khít với dữ liệu. Tức các luật có thể thực hiện tốt với dữ liệu huấn luyện, nhưng với dữ liệu sau đó thì kén đi. Để khắc phục, người ta xén luật. Luật được xén theo nghĩa bỏ các liên kết, tức các thử nghiệm thuộc tính. Luật R sẽ bị xén, nếu phiên bản đã xén R2 của nó có chất lượng tốt hơn, khi đánh giá trên tập các dòng độc lập.

Người ta sử dụng phương pháp hiệu quả theo FOIL. Đối với luật R, *hiệu quả xén* $(R) = (P - N) / (P + N)$. Giá trị này sẽ tăng theo độ chính xác của R trên tập đang xén. Nếu giá trị này càng cao đối với phiên bản đã xén của R, người ta xén R.

7.2.5. Đánh giá và lựa chọn mô hình

Sau khi xây dựng được mô hình phân lớp, người ta tự đặt ra vấn đề về (i) độ chính xác; (ii) độ tin cậy của mô hình.

7.2.5.1. Độ đo đánh giá hiệu năng của bộ phân lớp

Cần đánh giá bộ phân lớp tốt ra sao, tức chính xác ra sao, đối với việc dự đoán các nhãn lớp của các dòng dữ liệu. Khi các dòng của lớp ít nhiều phân tán, không cân đối, được xem như trường hợp phổ biến. Sau đây là một số khái niệm liên quan đến độ chính xác của mô hình phân lớp.

Công việc cần thực hiện là (i) dùng dữ liệu huấn luyện để xây dựng bộ phân lớp; (ii) đánh giá độ chính xác của mô hình đã học

được. Việc này có thể gặp tình trạng quá lí tưởng, do đặc biệt hóa thuật toán học trên dữ liệu. Để xác định đúng độ chính xác, người ta sử dụng *tập thử nghiệm*, gồm các dòng có nhãn lớp, mà chưa dùng đến trong lúc huấn luyện mô hình.

Liên quan đến thuật ngữ, người ta dùng khái niệm các *dòng dương*¹ đối với các dòng dữ liệu của lớp chính đang quan tâm; các dòng khác là *dòng âm*. Khi cho hai lớp, chẳng hạn dòng dương là *mua máy tính = có*; dòng âm là *mua máy tính = không*. Giả sử người ta sử dụng bộ phân lớp trên tập thử nghiệm với các dòng có nhãn. Số các dòng dương là P; âm là N.

Có bốn thuật ngữ nền, cho phép xây dựng các độ đo chính xác.

- *Dương đúng (TP²)*. TP cho biết các dòng dữ liệu dương, tức dòng có nhãn chấp thuận, được gán nhãn chính xác, tức gán nhãn đúng. Đây là trường hợp phân lớp *đúng*, gán nhãn *dương*;
- *Âm đúng (TN)*. TN cho biết các dòng âm, tức dòng mang nhãn từ chối, vẫn được bộ phân lớp gán nhãn chính xác, tức vẫn là âm. Đây là trường hợp phân lớp *đúng*, gán nhãn *âm*;
- *Dương sai (FP)*. FP là các dòng âm được phân lớp không chính xác, thành *dương*. Chẳng hạn các dòng đã có nhãn *mua máy tính = không*, tức âm, tức phân lớp *sai*, lại được cho là đúng. Đây là trường hợp phân lớp *sai*, gán cho dòng âm nhãn *dương*;
- *Âm sai (FN)*. FN là các dòng dương bị phân lớp sai, thành dòng nhãn âm. Đây là trường hợp phân lớp *sai*, gán cho dòng dương nhãn *âm*.

Hình sau trực quan hóa bốn thuật ngữ trên.

¹ Positive tuple, negative tuple: dòng dữ liệu dương, dòng dữ liệu âm.

² TP: true positive: dương đúng; TN: true negative: âm đúng; FP: false positive: dương sai; FN: false negative: âm sai.

Giả sử có hai lớp, có, không

Lớp mà bộ phân loại dự đoán

	Có	Không	Tổng	
Lớp thực sự của các dòng dữ liệu	Có	TP	FN	P
	Không	FP	TN	N
	Tổng	P'	N'	P + N

Hình. 8.12. Ma trận các thuật ngữ, còn được gọi là ma trận nhầm lẫn

Ma trận nhầm lẫn là công cụ tốt để phân tích bộ phân lớp tốt ra sao, khi nhận dạng các dòng của các lớp khác nhau. Bộ phân lớp đúng, với thông số TP và TN. Căn cứ vào thông số FP và FN, người ta biết mức độ phân lớp sai. Cho m lớp, $m > 1$, ma trận nhầm lẫn là bảng $m \times m$. Tại đầu vào $CM_{i,j}$ trong m dòng và m cột đầu tiên cho biết số các dòng của lớp i mà bị phân lớp thành lớp j . Đối với bộ phân lớp có độ chính xác tốt, lí tưởng là hầu hết các dòng được thể hiện theo đường chéo chính của ma trận, tức các giá trị $CM_{i,i}$ đến $CM_{m,m}$. Tức lí tưởng là FP, FN gần tới 0.

Trong bảng còn có giá trị P' và N' , là số các dòng được bộ phân lớp gán nhãn dương, âm. $P + N = P' + N'$ là tổng số các dòng dữ liệu. Trong chương trước đã trình bày một số khái niệm về độ chính xác. Đối với thuật toán phân lớp, nhắc lại chúng, với ngữ nghĩa của phân lớp.

1. *Độ chuẩn xác*, hay tỉ lệ nhận dạng đúng $(TP + TN) / (P + N)$. Thuật ngữ này liên quan đến việc so sánh với giá trị phân lớp với giá trị thực. Chẳng hạn khi bắn tên, người ta bắn quanh vòng 10, được xem là chuẩn xác;
2. *Tỉ lệ sai*, tức tỉ lệ phân lớp nhầm $(FP + FN) / (P + N)$;
3. *Độ chính xác*, chỉ sự lặp lại, liên quan đến giá trị đo, phân lớp. Chẳng hạn trong trò bắn tên, nếu không quanh vòng 10, nhưng các lần bắn đều chụm ở đâu đó, thì độ chính xác vẫn là cao. Độ chính xác là $TP / (TP + FP)$;

4. *Độ phủ mẫu đúng*, hay còn được gọi là độ nhạy¹. Độ đo này này cho biết tỉ lệ các dòng dương được phân lớp đúng, so với tổng số các dòng dương TP / P ;
5. *Độ phủ mẫu sai*², tức tỉ lệ các dòng âm được phân lớp đúng, đúng, so với tổng các dòng âm TN / N ;
6. *Trung bình điều hòa của độ chính xác và độ phủ mẫu đúng*. Gọi hai giá trị này là X, Y ; độ đo này xác định theo $2 * X * Y / (X + Y)$;
7. F_β Lấy β là số thực không âm, độ đo này được xác định $(1 + \beta^2) * X * Y / (\beta^2 * X + Y)$.

Bảng 7.2. Các độ đo đánh giá

Độ đo	Công thức
Chuẩn xác ³ , tỉ lệ nhận biết	$\frac{TP + TN}{P + N}$
Tỉ lệ sai, tỉ lệ phân lớp nhầm	$\frac{FP + FN}{P + N}$
Độ nhạy ⁴ , tỉ lệ dương đúng, hay độ phủ, hồi tưởng	$\frac{TP}{P}$
Đặc hiệu ⁵ , tỉ lệ âm đúng	$\frac{TN}{N}$
Chính xác ⁶	$\frac{TP}{TP + FP}$
F_1 , tỉ lệ f , trung bình điều hòa của độ chính xác (p) và độ nhạy (r)	$2 * \frac{p * r}{p + r}$
F_β , trong đó β là số thực không âm	$\frac{(1 + \beta^2) * p * r}{\beta^2 * p + r}$

¹ Recall: hồi tưởng; Sensitivity: nhạy.

² Specificity: đặc tính, đề nghị gọi là độ phủ mẫu sai.

³ Accuracy: chuẩn xác.

⁴ Sensitivity, recall: độ nhạy, độ phủ.

⁵ Specificity: đặc hiệu, cụ thể.

⁶ Precision: chính xác.

Trong bảng, các kí hiệu TP, TN, FP, P, N tham chiếu đến các số các mẫu *dương đúng*, tức tìm thấy đúng cái cần tìm, *âm đúng*, tức cái không tìm thấy nhưng đúng là cái cần tìm, *dương sai*, tức tìm thấy cái không cần tìm, *dương*, tức mẫu cần tìm, *âm*, tức mẫu không cần tìm.

Bảng 7.3. Ma trận nhầm lẫn đối với các lớp *mua máy tính = có* và *mua máy tính = không*, trong đó dòng *i* và cột *j* cho biết số bộ của lớp *i* bị bộ phân lớp đánh nhãn là *j*

Lớp	Mua máy tính = có	Mua máy tính = không	Tổng	Nhận dạng (%)
Mua máy tính = có	6954	46	7000	99.34
Mua máy tính = không	412	2588	3000	86.27
Tổng	7366	2634	10000	95.42

Trong bảng, các mục không trên đường chéo chính là không hay gần không, thì đó là trường hợp lí tưởng.

Về độ phủ mẫu đúng và độ phủ mẫu sai: hình cho thấy ma trận nhầm lẫn đối với dữ liệu *y* tế; có hai giá trị lớp đối với bệnh ung thư, với nhãn là *có*, *không*. Độ nhạy, hay độ phủ mẫu đúng, là $90 / 130 = 0.30$. Độ phủ mẫu sai $9560 / 9700 = 0.98$. Độ chính xác là $9650 / 10000 = 0.96$. Dù bộ phân lớp có độ chuẩn xác cao, khả năng nhận dạng đúng các mẫu dương là hạn chế. Người ta cần có kĩ thuật để quản lí các dữ liệu không cân đối này.

Bảng 7.4. Ma trận nhầm lẫn đối với các lớp *ung thư = mắc* và *ung thư = không*

Lớp	Mắc	Không	Tổng	Nhận dạng (%)
Mắc	90	210	300	30.00
Không	140	9560	9700	98.56
Tổng	230	9770	10000	96.40

Về độ chính xác và độ phù mẫu đúng: với dữ liệu trong bảng, độ chính xác đối với lớp có là $90 / 230 = 0.39$. Độ phù mẫu đúng là $90 / 300 = 0.30$.

Một vấn đề khác liên quan đến các độ đo chính xác của bộ phân lớp là các trường hợp mà người ta không nên sử dụng độ đo chính xác. Trong bài toán phân lớp, người ta thường giả thiết tất cả các dòng dữ liệu là phân lớp duy nhất, tức mỗi dòng huấn luyện chỉ thuộc về một lớp. Như vậy, trong cơ sở dữ liệu lớn, điều này không hợp lí. Thay vì vậy, nên giả thiết rằng *mỗi dòng có thể thuộc về nhiều lớp*. Khi đó, độ đo chuẩn xác không còn phù hợp.

Thay vì phân lớp nhân lớp, cần cho kết quả là *phân phối lớp xác suất*. Các độ đo chuẩn xác sử dụng kinh nghiệm, may rủi đoán thứ cấp¹, mà dự đoán về lớp được xem là đúng nếu nó khớp với lớp đầu tiên hay thứ hai có khả năng nhất. Mặc dù đây là một tiếp cận, nhưng nó chưa hoàn thiện.

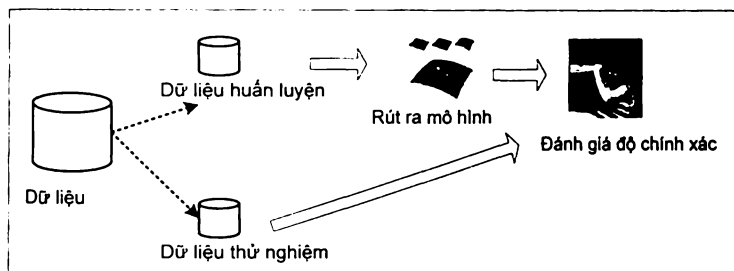
Ngoài các độ đo về chuẩn xác, các bộ phân lớp so sánh với nhau theo các khía cạnh khác:

- *Tốc độ*: liên quan đến giá tính toán để tạo và sử dụng bộ phân lớp;
- *Vững chắc*: đề cập khả năng của bộ phân lớp dự đoán đúng với các dữ liệu có ổn và thiếu;
- *Mở rộng*: liên quan đến khả năng áp dụng cho khối lượng lớn các dữ liệu, khi kích thước tăng lên;
- *Diễn giải*: hướng đến mức độ hiểu về bộ phân lớp. Thông số này là chủ quan, khó đánh giá. Cây quyết định và bộ phân lớp dùng luật là dễ hiểu, nhưng các bộ phân lớp khác là khó hiểu hơn.

¹ Second guess heuristic: may rủi đoán thứ cấp.

7.2.5.2. Phương pháp lấy ra và lấy mẫu ngẫu nhiên

Phương pháp *lấy ra*¹ liên quan đến đánh giá mức chuẩn xác. Theo phương pháp này, dữ liệu đã cho được chia ngẫu nhiên thành hai tập độc lập (i) *tập huấn luyện*; (ii) *tập thử nghiệm*. Diễn hình tập huấn luyện chiếm hai phần ba. Tập huấn luyện dùng để rút ra mô hình. Độ chuẩn xác của mô hình này được đánh giá với tập thử nghiệm. Việc đánh giá chưa hẳn là hay, hay bi quan, do chỉ dùng một phần của dữ liệu ban đầu để xây dựng mô hình.



Hình 8.13. Đánh giá độ chuẩn xác trong phương pháp lấy ra

Lấy mẫu ngẫu nhiên là biến thể của *phương pháp lấy ra*, khi phương pháp lấy ra được áp dụng k lần. Đánh giá độ chuẩn xác cuối cùng là trung bình của các độ chuẩn xác thành phần.

7.2.5.3. Kiểm chứng chéo

Trong *kiểm chứng chéo k lần*², dữ liệu khởi đầu được chia ngẫu nhiên thành k tập con loại trừ nhau, tức k khúc, $D_1, \dots, D_k$, có kích thước xấp xỉ nhau. Việc huấn luyện và thử nghiệm thực hiện k lần. Trong lần lặp thứ i , đoạn D_i dùng như tập thử nghiệm, còn các D_j khác hợp lại, tạo nên tập huấn luyện mô hình.

Không như các phương pháp lấy ra hay lấy mẫu ngẫu nhiên, phương pháp này được sử dụng cùng số lần để huấn luyện và thử nghiệm. Đối với phân lớp, đánh giá độ chuẩn xác là tỉ lệ giữa số

¹ Holdout: lấy ra.

² K - fold cross - validation: kiểm chứng chéo k lần.

các phân lớp đúng tổng thể trong k lần, so với tổng số các dòng dữ liệu trong tập ban đầu.

Phương pháp *lấy ra một*¹ là trường hợp đặc biệt của phương pháp kiểm chứng chéo k lần, với k là số các dòng ban đầu. Tức là chỉ một mẫu được lấy ra mỗi lần để thử nghiệm. Phương pháp kiểm chứng phân tầng² lấy các đoạn là phân tầng, sao cho phân phối lớp của các dòng trong mỗi đoạn là xấp xỉ nhau, như dữ liệu ban đầu. Nói chung, kiểm chứng chéo 10 lần là thông dụng trong đánh giá độ chuẩn xác, tùy vào độ lệch và sai khác tương đối thấp.

7.2.5.4. Khởi động lại

Phương pháp *khởi động lại*³ lấy mẫu trên các dòng huấn luyện theo thay thế. Mỗi lần, một dòng được chọn; dòng này vẫn có khả năng được chọn lại, do lại được trả về tập huấn luyện.

Có vài phương pháp khởi động lại. Phương pháp khởi động lại là thông dụng. Giả sử bộ dữ liệu có d dòng. Bộ dữ liệu được lấy mẫu d lần, với cách thay thế. Các dòng không tham gia huấn luyện sẽ được dùng làm dữ liệu thử nghiệm. Về trung bình, 0.632 các dòng được dùng cho mẫu khởi động, và 0.368 các dòng tạo nên tập thử nghiệm.

Về con số 0.632, người ta thấy mỗi dòng có xác suất chọn là $1/d$. Xác suất không được chọn là $1 - 1/d$. Người ta chọn d lần, nên xác suất không được chọn là $(1 - 1/d)^d$. Nếu d lớn, xác suất này xấp xỉ $e^{-1} = 0.368$.

Người ta có thể lặp lại thủ tục lấy mẫu k lần; trong mỗi lần lặp dùng tập thử nghiệm hiện có để thu được đánh giá về độ chuẩn xác của mô hình. Độ chuẩn xác chung của mô hình M được đánh giá theo $Acc(M) = \frac{1}{k} \sum_{i=1}^k (0.632 * Acc(M_{,1}) + 0.368 * Acc(M_{,2}))$;

¹ Leave one - out: lấy ra một.

² Stratified cross - validation: kiểm chứng chéo phân tầng.

³ Bootstrap: khởi động lại.

trong đó $\text{Acc}(M_1)_1$ là độ chuẩn xác của mô hình thu được với mẫu khởi động lại i , khi áp dụng trên bộ dữ liệu thử nghiệm i ; và $\text{Acc}(M_1)_2$ là trên bộ dữ liệu huấn luyện. Việc khởi động lại có xu hướng lí tưởng, tốt cho các bộ dữ liệu nhỏ.

7.2.5.5. Lựa chọn mô hình với các thử nghiệm thống kê về tính ý nghĩa

Giả sử tạo được hai mô hình phân lớp, M_1 và M_2 từ dữ liệu. Người ta thực hiện kiểm chứng chéo 10 lần để có tỉ lệ sai trung bình, tức $1 - \text{Acc}(M)$. Vấn đề là cần xác định mô hình tốt hơn. Về trực quan, người ta có thể chọn mô hình có tỉ lệ sai thấp hơn; tuy nhiên tỉ lệ sai trung bình chỉ *đánh giá* sai số trên lực lượng thực của các trường hợp dữ liệu tương lai. Có thể có sai khác đáng kể giữa các tỉ lệ sai trong bất kì thực nghiệm kiểm chứng chéo 10 lần. Dù các tỉ lệ sai trung bình ứng với M_1 , M_2 có thể khác nhau, nhưng sai khác không đủ có ý nghĩa. Do vậy cần đề cập cả những khác biệt khác.

Để xác định chúng sai khác thực sự không về tỉ lệ sai trung bình, người ta sử dụng *thử nghiệm về ý nghĩa thống kê*. Ngoài ra, người ta còn muốn có con số sai khác định lượng.

Về thực nghiệm thống kê, giả sử đối với mỗi mô hình, người ta dùng kiểm chứng chéo 10 lần; mỗi lần có phân đoạn dữ liệu khác nhau. Mỗi phân đoạn là độc lập. Người ta có thể tính trung bình các tỉ lệ sai đối với M_1 , M_2 . Đối với mô hình đã cho, các tỉ lệ sai riêng được tính theo kiểm chứng chéo được xem là mẫu độc lập, khác nhau từ phân phối xác suất. Nói chung, chúng theo *phân phối t với $k - 1$ độ tự do*; $k = 10$. Nó cho phép thực hiện thử nghiệm giả thuyết dùng trong *thử nghiệm t* , hay *thử nghiệm t Student*. Giả thuyết ở đây là hai mô hình như nhau; nếu bỏ giả thuyết này, hai mô hình sai khác về ý nghĩa, cho phép chọn mô hình có tỉ lệ sai thấp hơn.

Trong thực tế khai phá dữ liệu, người ta thường dùng tập thử nghiệm đơn, tức dùng chung bộ dữ liệu thử nghiệm đối với M_1 , M_2 . Khi đó người ta so sánh cặp hai mô hình đối với mỗi vòng kiểm

chúng 10 lần. Tức đối với vòng i của quá trình, sử dụng cùng một đoạn kiểm chứng chéo để xác định tỉ lệ sai cho M_1 , M_2 , được giá trị $e(M_1)_i$ và $e(M_2)_i$. Với M_1 có tỉ lệ sai trung bình $E(M_1)$; với M_2 có $E(M_2)$. Phương sai giữa hai mô hình là $v(M_1 - M_2)$. Thử nghiệm t tính thống kê t với $k - 1$ độ tự do đối với k mẫu. Thống kê t của so sánh cặp là $t = (E(M_1) - E(M_2)) / \sqrt{v(M_1 - M_2) / k}$; trong đó $v(M_1 - M_2) = \frac{1}{k} \sum_{i=1}^k (e(M_1) - e(M_2) - (E(M_1) - E(M_2)))^2$.

Để xác định sai khác có ý nghĩa giữa M_1 , M_2 , người ta tính t và chọn mức ý nghĩa, sig. Thực tế mức này là 5% hay 1%. Người ta tra bảng về phân phối t . Chẳng hạn muốn xác định sai khác giữa M_1 và M_2 là có ý nghĩa không đối với 95% lực lượng dữ liệu, tức sig = 5%; người ta tìm giá trị phân phối t ứng với $k - 1$ độ tự do. Tuy nhiên do phân phối t là đối xứng, trong bảng chỉ có giá trị phần trăm đầu; người ta tìm trong bảng giá trị $z = \text{sig} / 2$, tức 0.025. Tham số được xem là hạn chế tin cậy¹. Nếu $t > z$ hay $t < -z$, thì giá trị t thuộc vùng bị loại, trong phần đuôi của phân phối. Điều này có nghĩa người ta có thể bỏ giả thuyết về M_1 , M_2 như nhau, và khẳng định có sai khác có ý nghĩa. Ngược lại, nếu không bỏ giả thuyết, người ta cũng kết luận rằng bất kì sai khác giữa M_1 , M_2 đều có ý nghĩa.

Nếu dùng hai tập thử nghiệm, người ta sử dụng phiên bản khác của thử nghiệm t , trong đó phương sai giữa các tỉ lệ sai trung bình của hai mô hình là $v(M_1 - M_2) = \sqrt{v(M_1) / k_1 + v(M_2) / k_2}$; trong đó k_1 , k_2 là số các mẫu kiểm chứng chéo, dùng cho M_1 , M_2 . Đây được xem như thử nghiệm hai mẫu.

7.2.5.6. So sánh các bộ phân lớp theo lợi ích giá và theo đường cong đặc tính thực hiện

So sánh theo lợi ích giá

Các giá trị TP, TN, FP và FN có ích để đánh giá giá và lợi ích², hay rủi ro và lợi ích, liên quan đến mô hình phân lớp. Giá liên quan đến FN, chẳng hạn dự đoán sai bệnh nhân ung thư là không

¹ Confidence limit: hạn chế tin cậy.

² Cost and benefit: giá và lợi ích.

mắc bệnh. Trong những trường hợp này, người ta có thể tăng trọng số một loại sai sót so với sai sót khác, nhờ việc gán giá. Các giá này có thể là (i) nguy hiểm đối với bệnh nhân; (ii) giá trị liệu; (iii) chi phí viện phí. Tương tự, lợi ích liên quan đến quyết định TP có thể khác với TN. Việc tính độ chuẩn xác của bộ phân lớp hiện nay thường vẫn giả sử các giá đều nhau.

Theo tiếp cận khác, có thể kết hợp giá và lợi ích. Các ứng dụng cần đến phân tích giá - lợi ích gồm (i) quyết định vay; (ii) tiếp thị. Các giá khác được quan tâm trong phân tích tổng thể gồm (i) giá thu thập dữ liệu; (ii) giá phát triển công cụ phân lớp.

So sánh theo đường cong đặc tính thực hiện: đường đặc tính thực hiện¹ ROC là công cụ đồ họa thông dụng để so sánh hai mô hình phân lớp. Các đường ROC có nguồn gốc từ lý thuyết phát hiện tín hiệu, phát triển trong thế chiến II để phân tích hình ảnh radar. Đường ROC đối với mô hình đã cho biểu thị cân đối giữa tỉ lệ TP, kí hiệu TPR, và tỉ lệ FP, kí hiệu FPR². Cho tập thử nghiệm N dòng dữ liệu và một mô hình, $TPR = TP / P$, chính là độ phủ mẫu đúng; $FPR = FP / N$, chính là 1 - độ phủ mẫu sai.

Đối với bài toán có hai lớp, đường ROC cho phép hiển thị cân đối giữa tỉ lệ mà mô hình có thể nhận dạng chuẩn xác các trường hợp dương, so với tỉ lệ mà mô hình nhận dạng nhầm các trường hợp âm như là dương, đối với các phần khác nhau của tập thử nghiệm. Bất kì gia tăng trong TPR liên quan đến giá của việc tăng về FPR. Miền dưới của đường ROC là độ đo về chuẩn xác của mô hình.

Để vẽ đường ROC cho mô hình phân lớp M, mô hình này cần cho biết xác suất của lớp dự đoán đối với mỗi dòng thử nghiệm. Với thông tin này, người ta xếp hạng và sắp xếp các dòng, để dòng dữ liệu có nhiều khả năng thuộc về lớp dương, tức có, hiện ở đầu

¹ ROC: curve, Receiver Operating Characteristic curve: đường đặc tính thực hiện.

² TPR, FPR: True Positive Rate, False Positive Rate: tỉ lệ dương đúng, tỉ lệ dương sai.

danh sách; và dòng dữ liệu ít có khả năng thuộc vào lớp dương sẽ ở cuối danh sách. Các bộ phân lớp Bayes đơn giản và bộ phân lớp truyền ngược cho biết phân phối xác suất về lớp đối với mỗi dự đoán, là thích hợp. Các bộ phân lớp khác, như cây quyết định, có thể dễ điều chỉnh để cho biết dự đoán về xác suất lớp lấy giá trị mà bộ phân lớp trả về đối với dòng dữ liệu X là $f(X) \rightarrow [0, 1]$. Đối với bài toán nhị phân, ngưỡng t được chọn để các dòng có $f(X) \geq t$ được coi như dương, và các dòng khác là âm. Số các TP và FP đều là hàm của t , sao cho có thể viết $TP(t)$, $FP(t)$. Cả hai là hàm đơn điệu giảm.

Ý tưởng chung của việc vạch đường ROC: trục tung của ROC thể hiện TPR; trục hoành thể hiện FPR. Để vẽ đường ROC, trước tiên xuất phát từ góc dưới trái, tức $TPR = FPR = 0$, rồi kiểm tra nhãn lớp thực sự của dòng dữ liệu tại đầu danh sách. Nếu có TP, thì tăng TP, tức tăng cả TPR. Trên đồ thị, người ta di chuyển lên và chấm một điểm. Ngược lại, nếu bộ phân lớp dự đoán dòng âm như dương, tức có FP, thì FP và FPR tăng. Trên đồ thị, người ta di chuyển sang phải và chấm một điểm. Quá trình này lặp lại đối với mỗi dòng thử nghiệm theo thứ tự đã xếp hạng; mỗi lần di chuyển lên trên đồ thị đối với TP hay sang phải đối với FP.

Vẽ về đường ROC: hình cho thấy các giá trị xác suất do bộ phân lớp xác suất trả về đối với 10 dòng dữ liệu của tập thử nghiệm. Các dòng được xếp thứ tự giảm dần theo xác suất. Các thông tin gồm (i) tên dòng; (ii) lớp thực sự của dữ liệu. Trong bảng có $P = 5$, $N = 5$.

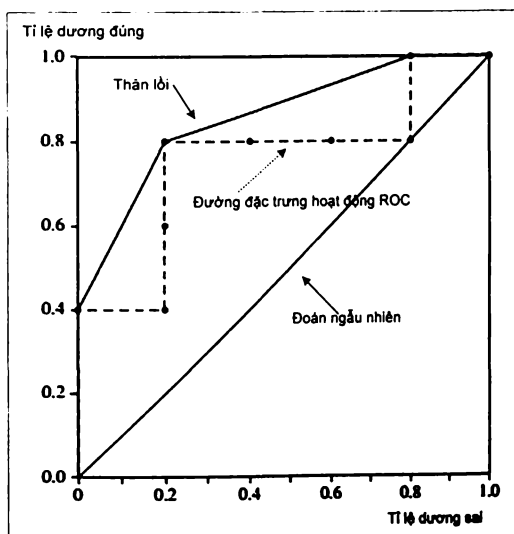
Bắt đầu từ dòng 1, xác suất cao nhất, có điểm số được lấy làm ngưỡng, $t = 0.9$. Do đây là dương đúng, $TP = 1$, $FP = 0$. Do vậy $TN = 5$, $FN = 4$. từ đó tính được $TPR = TP / P = 1 / 5 = 0.2$, $FPR = 0$. Vậy người ta có điểm $(0.2, 0)$ trên đường ROC.

Tiếp tục với dòng thứ hai, t được lấy là 0.8, ứng với xác suất dòng này. Đây là dòng dương, nên $TP = 2$, tương tự, người ta có điểm $(0.4, 0)$ trên đường ROC.

Với dòng thứ ba, $t = 0.7$. Dòng này được coi là dương, trong khi nó là âm, nên nó là FP. Vậy TP không đổi, còn FP tăng, FP = 1. Người ta có điểm (0.4, 0.2) trên đường ROC.

Bảng 7.5. Các dòng dữ liệu sắp xếp theo điểm giảm dần; trong đó điểm là giá trị đo bộ phân lớp xác suất trả về

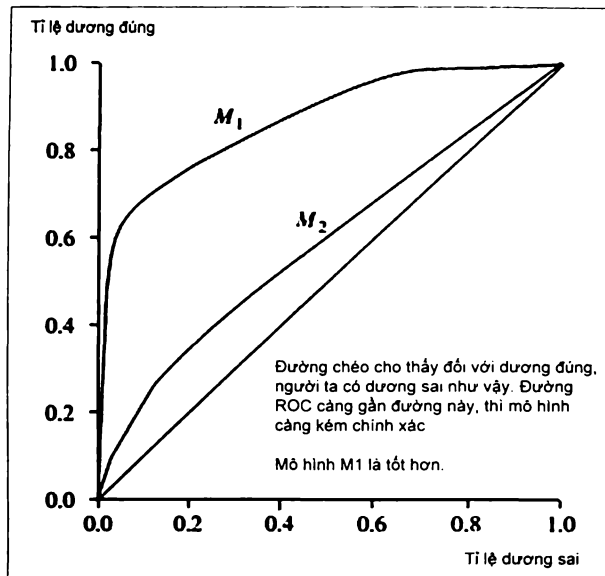
Số bộ	Lớp	Xác suất	TP	FP	TN	FN	Tỉ lệ dương đúng TPR	Tỉ lệ dương sai FPR
1	P	0.90	1	0	5	4	0.2	0
2	P	0.80	2	0	5	5	0.4	0
3	N	0.70	2	1	4	3	0.4	0.2
4	P	0.60	3	1	4	2	0.6	0.2
5	P	0.55	4	1	4	1	0.8	0.2
6	N	0.54	4	2	3	1	0.8	0.4
7	N	0.53	4	3	2	1	0.8	0.6
8	N	0.51	4	4	1	1	0.8	0.8
9	P	0.50	5	4	0	1	1.0	0.8
10	N	0.40	5	5	0	0	1.0	1.0



Hình. 7.14. Đường cong ROC đối với dữ liệu trong thí dụ

Trong lý thuyết phát hiện tín hiệu, *đặc trưng hoạt động của bộ thu ROC¹*, hay *đường cong đặc trưng hoạt động của bộ thu*, là đồ thị trong hệ thống phân lớp nhị phân.

Có nhiều phương pháp để có đồ thị ngoài các điểm này; phương pháp thông dụng nhất dùng bao lồi. Đồ thị cho thấy đường chéo nối các điểm TP của mô hình. Để so sánh, đường này thể hiện dự đoán ngẫu nhiên.



Hình 7.15. Đường ROC đối với hai mô hình

Hình cho thấy đường ROC của hai mô hình phân lớp. Đường chéo thể hiện dự đoán ngẫu nhiên. Vậy đường ROC nào càng gần đường chéo thì càng kém chuẩn xác. Nếu mô hình thực sự tốt, ban đầu thường từ các TP, khi duyệt từ trên xuống trong danh sách. Điều này có nghĩa đường ROC đi dần lên từ gốc tọa độ. Về sau, người ta gặp ít dần các TP, gặp FP nhiều hơn; đường ROC sẽ đi ngang.

¹ ROC: receiver operating characteristic: đặc trưng hoạt động của bộ thu;
ROC: receiver operating curve: đường cong hoạt động của bộ thu.

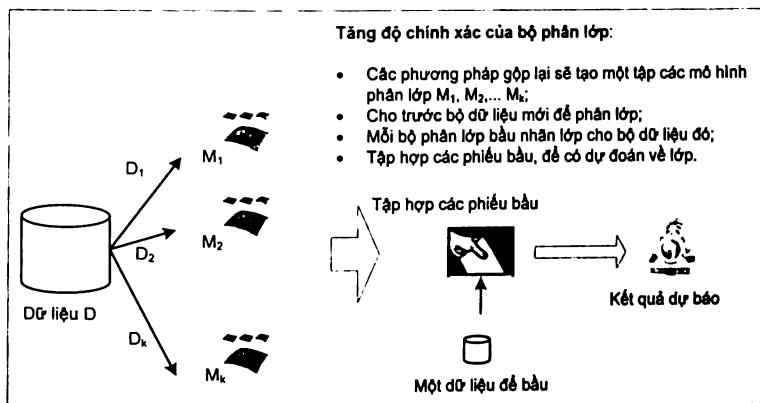
7.2.6. Các kỹ thuật cải thiện độ chuẩn xác phân lớp

Người ta đề xuất các *phương pháp tập thể*¹, tức kết hợp, tập hợp, nhiều mô hình phân lớp. Có ba phương pháp thuộc loại này (i) gộp các phương pháp khởi động lại; (ii) phương pháp lấy trọng số; (iii) phương pháp tập các cây quyết định.

Ngoài ra cần giải quyết vấn đề về dữ liệu có phân phối mất cân đối. Việc này khắc phục các mô hình học truyền thống.

7.2.6.1. Về các phương pháp tập thể

Hình cho thấy hoạt động của các phương pháp phân lớp tập thể. Người ta dùng tập k mô hình học, $M_1, \dots, M_k$, để tạo mô hình phân lớp gộp M^* . Cho bộ dữ liệu D , dùng để tạo k tập huấn luyện $D_1, \dots, D_k$, trong đó D_1 được dùng để tạo nên bộ phân lớp M_1 . Khi cho một dòng dữ liệu cần phân lớp, các bộ phân lớp cơ bản sẽ bình bầu để cho kết quả là lớp của dòng dữ liệu này. Tập thể cho ra dự đoán về lớp, dựa vào bầu chọn giữa các bộ phân lớp cơ bản.

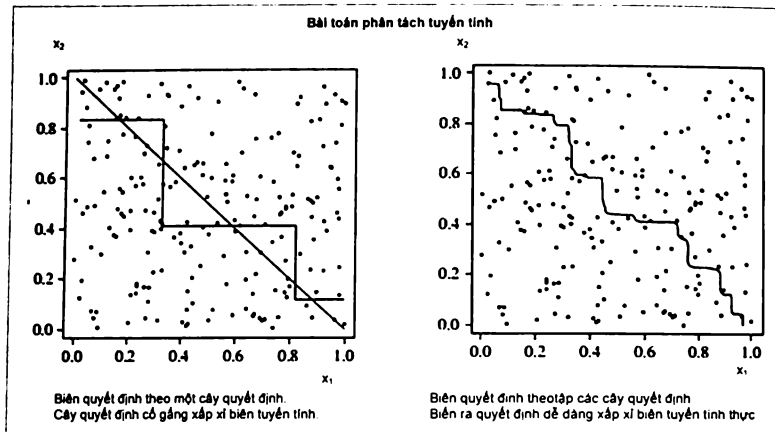


Hình 7.16. Dùng tập thể nhiều bộ phân lớp để tăng độ chuẩn xác

Xét bài toán đơn giản với hai lớp, mô tả qua hai thuộc tính x_1, x_2 . Hình thể hiện biên quyết định theo bộ phân lớp cây quyết định; và

¹ Ensemble method: phương pháp tập thể.

biên quyết định với tập thể các cây quyết định. Trường hợp sau cho kết quả mịn hơn.



Hình 7.17. Biên phân lớp nhờ cây quyết định

7.2.6.2. Phương pháp gộp các phương pháp khởi động lại

Xuất phát của các phương pháp tập thể để tăng độ chuẩn xác có thể hình dung như người bệnh mắc trọng bệnh, muốn đi khám k bác sĩ, rồi lựa chọn phân lớp bệnh theo cơ chế đa số.

Cho tập D có d dòng, phương pháp gộp các phương pháp khởi động lại¹ tiến hành k lần lặp i với tập huấn luyện D_i có d dòng; d dòng này được lấy mẫu theo cách thay thế từ bộ dữ liệu gốc D . Mỗi tập huấn luyện là một mẫu lấy lại; một vài dòng ban đầu trong D có thể không có mặt trong D_i , và cũng có dòng xuất hiện nhiều lần. Để phân lớp dòng X chưa biết, mỗi bộ phân lớp M_i cho biết một kết quả dự đoán, được tính như một phiếu bầu. Bộ phân lớp gộp M^* tính số phiếu, và gán lớp có số phiếu tối đa cho dòng dữ liệu X này. Phương pháp này có thể dùng cho các giá trị liên tục, nhờ việc lấy giá trị trung bình của mỗi dự đoán đối với dòng dữ liệu thử nghiệm đã cho.

¹ Bagging: túi, sử dụng kĩ thuật như gộp các phương pháp khởi động lại trong việc chọn tập thử nghiệm. Thuật ngữ này có nghĩa *Bootstrap aggregation*.

Bộ phân lớp gộp các phương pháp khởi động lại có độ chuẩn xác cao hơn bộ phân lớp đơn, trên cùng bộ dữ liệu D . Nó cũng vững chắc đối với dữ liệu có ồn hay bị thiếu.

Thuật toán ứng với phương pháp gộp các phương pháp khởi động lại cho phép tạo tập các mô hình phân lớp đối với lược đồ học. Mỗi mô hình đều bình đẳng về phiếu dự đoán, tức cùng trọng số.

Đầu vào: (i) tập các dòng huấn luyện D ; (ii) số các mô hình k ; (iii) lược đồ học phân lớp, chẳng hạn thuật toán cây quyết định, hay Bayes đơn giản...

Đầu ra: Mô hình tập thể M^* .

Thuật toán:

For $i = 1$ to k do // tạo k mô hình

 Tạo mẫu khởi động lại, D_i , nhờ lấy mẫu thay thế trên D ;

 Từ D_i và lược đồ học, rút ra mô hình M_i ;

Endfor

7.2.6.3. Phương pháp trọng số và trọng số phù hợp

Lại nói về bệnh nhân đi khám k bác sĩ với cùng triệu chứng, với hi vọng chọn được kết quả dự đoán chuẩn xác nhất. Trong phương pháp trọng số, kết quả khám của bác sĩ i có trọng số khác với bác sĩ j .

Trong *phương pháp trọng số*¹, người ta gán các trọng số cho các dòng huấn luyện dãy k bộ phân lớp tiến hành học lặp lại. Sau một lần phân lớp thứ i nhờ M_i , các trọng số được cập nhật, cho phép các bộ phân lớp sau $M_{i+1}, \dots$ lưu ý hơn đến các dòng mà M_i phân lớp nhầm. Bộ phân lớp cuối cùng M^* kết hợp các phiếu bầu của mỗi bộ phân lớp, với trọng số của mỗi phiếu là hàm của độ chuẩn xác của nó.

¹ Boosting: thúc đẩy, theo trọng số.

*Phương pháp trọng số thích nghi*¹ là thuật toán trọng số thông dụng. Giả sử người ta muốn lấy trọng số cho độ chuẩn xác của phương pháp học. Có bộ dữ liệu D với d dòng có nhãn (X_i, Y_i) ; trong đó y_i là nhãn lớp của dòng X_i . Ban đầu phương pháp gán cho mỗi dòng huấn luyện trọng số đều $1/d$. Người ta qua k vòng để tạo được k bộ phân lớp. Tại vòng thứ i , các dòng từ D được lấy mẫu để tạo tập huấn luyện D_i , kích thước d . Lấy mẫu theo kỹ thuật thay thế. Mỗi dòng sẽ được chọn theo trọng số của nó. Mô hình phân lớp M_i thu được từ bộ dữ liệu huấn luyện D_i . Sai số của nó được tính theo D_i , như tập thử nghiệm. Các trọng số của các dòng huấn luyện được điều chỉnh theo cách chúng được phân lớp.

Nếu một dòng bị phân lớp không chính xác, trọng số của nó tăng lên. Nếu dòng được phân lớp đúng, trọng số sẽ giảm. Trọng số của dòng dữ liệu phản ánh mức khó khi phân lớp. Trọng số càng cao thì nó thường bị phân lớp sai. Để tính tỉ lệ sai của mô hình M_i , người ta lấy tổng của mỗi dòng trong D_i mà M_i nhầm lẫn, tức $e(M_i) = \sum_{j=1}^d w_j * e(X_j)$; trong đó $e(X_i)$ là sai do phân lớp nhầm của X_i . Nếu nó bị phân lớp sai, $e(X_i) = 1$; ngược lại $e(X_i) = 0$. Nếu hiệu năng của bộ phân lớp M_i kém, mà sai số $e(M_i) > 0.5$, người ta có thể loại bỏ M_i . Thay vào đó, người ta tạo M_i mới từ bộ huấn luyện D_i mới.

Một khi đã xác định được trọng số cho các bộ phân lớp, người ta đề cập việc dự đoán nhãn cho dòng X . Không như tập thể các phương pháp khởi động lại với phiếu bầu bình đẳng, ở đây mỗi bộ phân lớp có trọng số đối với phiếu bầu, căn cứ vào hiệu quả phân lớp của bộ phân lớp. Tỉ lệ sai càng thấp, tức bộ phân lớp càng chính xác, thì bộ phân lớp có trọng số cao cho phiếu bầu. Trọng số của phiếu bầu của bộ phân lớp M_i là $\log((1 - e(M_i)) / e(M_i))$. Đối với mỗi lớp C , người ta lấy tổng các trọng số của mỗi bộ phân lớp đã gán lớp C cho X .

Liên quan đến so sánh phương pháp trọng số với tập thể các phương pháp khởi động lại, người ta thấy do cách lấy trọng số tập

¹ AdaBoost Adaptive Boosting: trọng số thích nghi.

trung vào các dòng bị nhầm, có nguy cơ quá khít trong mô hình tập thể.

Đôi khi kết quả của mô hình theo trọng số có thể kém chuẩn xác so với mô hình đơn trên cùng bộ dữ liệu. Tập thể các phương pháp khởi động lại ít nhạy với việc quá khít. Khi kết hợp các bộ phân lớp thành phần, phương pháp lấy trọng số có xu hướng đạt độ chuẩn xác cao hơn.

Thuật toán lấy trọng số thích nghi tạo nên tập các bộ phân lớp; mỗi bộ có một trọng số cho thiếu bấu.

Đầu vào: (i) bộ dữ liệu D ; (ii) số vòng lặp k , tức mỗi lần phân lớp ứng với một vòng lặp lại; (iii) lược đồ học phân lớp.

Thuật toán:

Lấy trọng số cho mỗi dòng dữ liệu là $1 / d$;

For $i = 1$ to k do // thực hiện mỗi vòng lặp

Lấy mẫu D bằng kĩ thuật thay thế, theo trọng số của dòng, thu được D_i ;

Dùng tập huấn luyện D_i để rút ra mô hình M_i ;

Tính sai số mô hình $e(M_i)$;

If $e(M_i) > 0.5$ then

Quay lại bước lấy mẫu, để thực hiện lại việc lấy mẫu D_i ;

EndIf

For mỗi dòng t trong D_i được phân lớp đúng do

Trọng số $(t) = \text{trọng số}(t) * (e(M_i) / (1 - e(M_i)))$; // cập nhật trọng số

Endfor

Thuật toán phân lớp dòng X nhờ tập thể các bộ phân lớp:

Thiết lập trọng số đối với mỗi lớp là 0;

For $i = 1$ to k do // đối với mỗi bộ phân lớp

$w_i = \log ((1 - e^{-M_i}) / e^{-M_i});$ // trọng số của phiếu bầu

$C = M_i(X);$ // lấy dự đoán lớp cho X theo mô hình M_i

Bổ sung w_i vào trọng số của lớp C ;

Endfor

Trả về lớp có trọng số lớn nhất;

7.2.6.4. Phương pháp tập các cây quyết định

Phương pháp sử dụng nhiều bộ phân lớp cây quyết định, hay *rừng ngẫu nhiên*¹, tức *tập các cây quyết định*. Cây quyết định riêng lẻ có được do lấy ngẫu nhiên các thuộc tính tại nút, để xác định phân tách. Tức mỗi cây phụ thuộc vào các giá trị của vectơ ngẫu nhiên, được lấy mẫu độc lập, với cùng phân phối. Các phiếu bầu của cây và lớp thông dụng nhất là kết quả phân lớp.

Phương pháp tập các cây quyết định có thể được xây dựng theo cách khởi động lại, kết hợp với lựa chọn thuộc tính. Với bộ dữ liệu D có d dòng, thủ tục chung để tạo k cây quyết định: đối với mỗi $i = 1..k$, lấy mẫu tập D_i có d dòng theo kỹ thuật thay thế từ D . F là số các thuộc tính cho phép xác định việc tách cây; F nhỏ hơn số các thuộc tính. Để kiến thiết cây quyết định M_i , tại mỗi nút, người ta chọn ngẫu nhiên F thuộc tính để tách cây tại nút đó. Dùng phương pháp luận CART để phát triển cây. Các cây này có kích thước lớn và không bị xén. Rừng ngẫu nhiên được xây dựng, theo *lựa chọn đầu vào ngẫu nhiên*, được gọi là *rừng đầu vào ngẫu nhiên*, *rừng RI^2* .

Một dạng khác là *rừng RC^3* , dùng tổ hợp tuyến tính ngẫu nhiên của các thuộc tính đầu vào. Thay vì chọn tập con các thuật toán một cách ngẫu nhiên, phương pháp này tạo các thuộc tính

¹ Random Forest: rừng ngẫu nhiên, chỉ tập thể các phương pháp cây quyết định.

² Forest with random input selection: rừng RI .

³ Forest with random linear combination: rừng RC .

mới, tức các đặc trưng, theo tổ hợp tuyến tính của các thuộc tính đã có. Thuộc tính mới được tạo nên nhờ bổ sung thuộc tính với hệ số, chính là các số ngẫu nhiên chuẩn thuộc $[-1, 1]$. Người ta tạo nên F tổ hợp tuyến tính, và thực hiện tách cây theo chúng. Việc này tạo nên rừng ngẫu nhiên; nó có ích khi chỉ có ít thuộc tính, nên giảm được tương quan giữa các bộ phân lớp riêng lẻ.

Phương pháp rừng ngẫu nhiên vững chắc đối với dữ liệu lẻ và lỗi, về độ chuẩn xác, so với phương pháp trọng số thích nghi. Sai số chung của rừng là tổng hợp từ các sai số trên cây, nên lớn. Độ chuẩn xác của rừng ngẫu nhiên tùy thuộc vào các bộ phân lớp riêng lẻ. Phương pháp này nhạy cảm với số các thuộc tính được chọn ở mỗi lần tách cây. Diễn hình lựa chọn $\log_2 d + 1$.

7.2.6.5. Cải thiện độ chuẩn xác phân lớp đối với dữ liệu không cân đối về lớp

Cho dữ liệu hai lớp. Phần này đề cập *vấn đề mất cân bằng về lớp*.

Định nghĩa: Dữ liệu được gọi là không cân đối theo lớp nếu lớp quan tâm chính, tức lớp dương, được thể hiện cho qua một ít các dòng, trong khi phân lớp các dòng thể hiện lớp âm.

Đối với dữ liệu mất cân đối theo nhiều lớp, phân phối dữ liệu của mỗi lớp khác nhau, nhưng lớp chính có phân phối hiếm hoi. Vấn đề mất cân đối về lớp liên quan chặt chẽ với việc học nhạy cảm với giá, mà chi phí cho sai số theo các lớp là khác nhau. Trong chẩn đoán y khoa, nguy hiểm về chẩn đoán người bệnh ung thư là khỏe, tức âm sai, có lỗi nặng hơn so với chẩn đoán người lành mắc bệnh ung thư, tức dương sai. Các ứng dụng khác gặp hiện tượng mất cân đối về lớp là (i) phát hiện lỗi; (ii) phát hiện tràn dầu trên ảnh vệ tinh; (iii) giám sát lỗi.

Các thuật toán phân lớp truyền thống nhằm tối thiểu hóa số sai sót trong quá trình phân lớp. Người ta giả sử giá của FP và FN là đều nhau. Do giả thiết phân phối cân đối về lớp và về giá sai sót, chúng không phù hợp đối với các dữ liệu mất cân đối về lớp.

Các tiếp cận tổng quát để cải thiện độ chuẩn xác phân lớp của dữ liệu mất cân đối là (i) lấy mẫu nhiều hơn; (ii) lấy mẫu ít đi; (iii) di chuyển ngưỡng; (iv) các kĩ thuật phân lớp tập thể. Ba tiếp cận đầu không đòi hỏi bất kì thay đổi nào trên kiến trúc của mô hình phân lớp. Tức việc lấy mẫu nhiều hay ít đi thay đổi phân phối của các dòng trong tập huấn luyện. Di chuyển ngưỡng tác động đến cách mô hình ra quyết định khi phân lớp dữ liệu mới.

Việc *lấy mẫu nhiều lên* lấy mẫu nữa với các dòng dương, để có bộ dữ liệu huấn luyện cân bằng giữa dòng dương và dòng âm. *Lấy mẫu ít đi* giảm số dòng âm; nó loại bỏ ngẫu nhiên các dòng từ lớp đa số, âm, cho đến khi các dòng cân bằng.

Về lấy mẫu nhiều lên, ít đi: giả sử bộ dữ liệu huấn luyện có 100 dòng dương và 1000 dòng âm. Đối với cách *lấy mẫu nhiều lên*, người ta nhân lên các dòng của lớp hiếm, để tạo bộ dữ liệu mới với 1000 dòng dương và 1000 dòng âm. Đối với cách *lấy mẫu ít đi*, người ta loại bỏ ngẫu nhiên các dòng âm, thu được tập huấn luyện mới với 100 dòng dương và 100 dòng âm.

Tiếp cận *di chuyển ngưỡng* đối với vấn đề mất cân bằng lớp không yêu cầu lấy mẫu. Cho dòng dữ liệu đầu vào, bộ phân lớp cho ra kết quả, $f(X) \rightarrow [0, 1]$. Thay vì xử lí các dòng huấn luyện, phương pháp này trả về quyết định phân lớp dựa trên đầu ra. Theo cách đơn giản nhất, với ngưỡng t , các dòng có $f(X) > t$ được xem là dương; các dòng khác là âm. Các tiếp cận khác đòi hỏi xử lí đầu ra với trọng số. Nói chung, ngưỡng dịch chuyển về t , để các dòng lớp hiếm dễ được phân lớp; vì vậy ít cơ hội bị tổn kém do sai số âm sai. Phương pháp dịch chuyển ngưỡng không thông dụng như việc lấy mẫu nhiều lên, ít đi, nhưng nó đơn giản và hiệu quả đối với vấn đề dữ liệu mất cân đối hai lớp.

Các phương pháp tập thể cũng ứng dụng được cho vấn đề mất cân bằng lớp. Các bộ phân lớp riêng lẻ thực hiện theo ba cách vừa nêu trên.

Các phương pháp này làm việc tương đối tốt đối với bài toán mất cân đối hai lớp. Kinh nghiệm cho thấy các phương pháp tập

thể và dịch chuyển ngưỡng tốt hơn phương pháp lấy mẫu lại. Phương pháp dịch chuyển ngưỡng làm việc tốt với bộ dữ liệu thậm chí rất mất cân đối. Bài toán mất cân đối nhiều lớp là phức tạp hơn, mà phương pháp lấy mẫu lại cho thấy kém hiệu quả.

7.2.7. Kết luận về phân lớp

Một số khía cạnh đã trình bày trong chương này, liên quan đến phân lớp:

- *Phân lớp* là dạng phân tích dữ liệu, cho phép trích ra mô hình mô tả các lớp dữ liệu. Các mô hình *dự đoán* số mô hình hóa các hàm giá trị liên tục. Dự đoán số và phân lớp là hai dạng chính của bài toán dự đoán;
- *Quy nạp cây quyết định* là thuật toán quy nạp đệ quy trên cây, sử dụng độ đo chọn thuộc tính để lấy ra thuộc tính, cho phép thử đối với mỗi nút lá trong cây. Các thuật toán *ID3*, *C4.5* và *CART* dùng các độ đo chọn thuộc tính khác nhau. Các thuật toán xén cây cải thiện độ chuẩn xác nhờ loại bỏ các nhánh cây ứng với ồn dữ liệu. Các thuật toán cây quyết định thế hệ đầu giả sử dữ liệu đặt trong bộ nhớ. Có thuật toán mở rộng dùng để quy nạp cây mở rộng;
- *Phân lớp Bayes đơn giản* dựa vào định lý Bayes về xác suất tiên nghiệm. Nó giả thiết có độc lập điều kiện về lớp, tức tác động của giá trị thuộc tính trên lớp đã biết là độc lập với các giá trị của thuộc tính khác;
- *Bộ phân lớp dùng luật* có tập các luật IFTHEN. Luật có thể được trích từ cây quyết định. Người ta có thể tạo trực tiếp các luật từ dữ liệu huấn luyện, nhờ thuật toán phù tuần tự;
- *Ma trận nhầm lẫn* dùng để đánh giá chất lượng của bộ phân lớp. Đối với bài toán hai lớp, có các khái niệm TP, TN, FP, FN. Các độ đo đánh giá khả năng dự đoán của bộ phân lớp là độ chuẩn xác, độ phù mẫu đúng, hay nhạy, độ

phù mẫu sai, độ chính xác, F và F_B. Người ta không tin cậy về độ chuẩn xác khi lớp quan tâm chính có ít dòng dữ liệu;

- Kiến thiết và đánh giá bộ phân lớp đòi hỏi phân đoạn dữ liệu có nhãn thành (i) tập huấn luyện; (ii) tập thử nghiệm. Các phương pháp phân đoạn điển hình là (i) lấy ra; (ii) lấy mẫu ngẫu nhiên; (iii) kiểm chứng chéo; (iv) khởi động lại;
- Thử về ý nghĩa và đường ROC là các công cụ để lựa chọn mô hình. *Thử tính có ý nghĩa* dùng để đánh giá sự xuất hiện về sai khác giữa hai mô hình phân lớp theo cơ hội. Đường ROC vạch tỉ lệ TP, hay độ phù mẫu đúng, so với tỉ lệ TN, tức 1 - độ phù mẫu sai, của một hay nhiều bộ phân lớp;
- Các phương pháp tập thể dùng để tăng độ chuẩn xác tổng thể nhờ học và kết hợp các mô hình phân lớp riêng lẻ. Các phương pháp tập thể thông dụng là (i) tập thể các phương pháp khởi động lại; (ii) trọng số; (iii) tập các cây quyết định;
- Vấn đề mất cân đối về lớp xảy ra khi lớp quan tâm chính chỉ thể hiện qua một số ít các dòng dữ liệu. Các chiến lược cho vấn đề này là (i) lấy mẫu nhiều lên; (ii) lấy mẫu ít đi; (iii) di chuyển ngưỡng; (iv) các kĩ thuật tập thể.

Ngoài những khía cạnh cơ bản về phân lớp dữ liệu đã trình bày trong chương này, người ta có thể phát triển các khía cạnh cao hơn:

1. Các mạng tin cậy theo xác suất. Nội dung này đề cập (i) các khái niệm và cơ chế về mạng tin cậy; (ii) huấn luyện trên mạng tin cậy Bayes;
2. Phân lớp theo cách truyền ngược. Nội dung này đề cập (i) mạng nơron truyền thẳng nhiều lớp; (ii) cấu hình mạng; (iii) truyền ngược; (iv) diễn giải việc truyền ngược;
3. Các máy vector tựa. Nội dung này đề cập (i) trường hợp dữ liệu tách biệt tuyến tính; (ii) trường hợp dữ liệu không tách biệt tuyến tính;

4. Phân lớp với các *mẫu thường xuyên*. Nội dung này đề cập (i) phân lớp kết hợp; (ii) phân lớp dựa trên mẫu thường xuyên phân biệt;
5. *Học từ bên cạnh*. Nội dung này đề cập (i) bộ phân lớp k láng giềng gần nhất; (ii) lập luận theo trường hợp;
6. Một số bộ phân lớp khác, sử dụng (i) thuật toán di truyền; (ii) theo tiếp cận tập thô; (iii) theo tiếp cận tập mờ;
7. Một số khía cạnh liên quan đến phân lớp (i) phân lớp nhiều lớp; (ii) phân lớp nửa giám sát; (iii) học chủ động; (iv) học chuyển giao.

7.3. Phân tích cụm

7.3.1. Bài toán phân cụm

Hình dung rằng giám đốc của *Hàng điện tử* muốn có năm người quản lí, và muốn các khách hàng được tổ chức thành năm nhóm, sao cho mỗi nhóm gần với người quản lí khác nhau. Vấn đề là hai khách hàng với mẫu kinh doanh khác nhau không vào cùng nhóm. Bởi vậy cần phát triển quan hệ khách hàng cho mỗi nhóm, dựa trên đặc trưng chung của các khách hàng trong nhóm.

Trong bài toán phân lớp, người ta biết nhãn lớp của mỗi khách hàng. Ở đây chưa biết nhãn lớp của khách hàng, mà cần *phát hiện* việc chia nhóm. Vậy bài toán là với số lớn các khách hàng và nhiều thuộc tính mô tả đặc điểm khách hàng, người ta cần nghiên cứu về dữ liệu và có cách phân hoạch khách hàng thành các nhóm. Việc này có thể tốn kém và kém khả thi.

*Định nghĩa: Phân cụm*¹ là quá trình chia nhóm tập các đối tượng dữ liệu thành nhiều nhóm, hay cụm, sao cho các đối tượng trong một cụm tương tự nhau cao, nhưng tương tự thấp đối với các đối tượng nhóm khác.

¹ Clustering: phân cụm.

Tính tương tự và khác biệt¹ được đánh giá dựa trên các giá trị thuộc tính mô tả đối tượng, thường liên quan đến độ đo khoảng cách.

Phần này trình bày khái niệm về phân cụm, các mô hình phân cụm và các kĩ thuật phân cụm cơ bản.

7.3.1.1. Khái niệm về phân tích cụm

Phân tích cụm, hay *phân cụm*, là quá trình phân hoạch tập các đối tượng dữ liệu, hay các quan sát, thành các tập con. Mỗi tập con là một *cụm*, mà các đối tượng trong cụm là tương tự nhau, nhưng các đối tượng này cần khác biệt với các đối tượng trong cụm khác. Việc phân cụm thu được tập các cụm, do quá trình phân tích cụm. Phương pháp phân cụm khác nhau sẽ thu được kết quả phân cụm khác nhau, trên cùng bộ dữ liệu. Việc *phân hoạch*² do thuật toán phân cụm thực hiện, chứ không thực hiện thủ công.

Ứng dụng của phân tích cụm có nhiều (i) thông minh doanh nghiệp; (ii) nhận dạng; (iii) tìm kiếm Web; (iv) sinh học; (v) an toàn dữ liệu. Chẳng hạn trong thông minh doanh nghiệp, người ta dùng phân cụm để tổ chức số lượng lớn các khách hàng thành nhóm, cùng chung đặc tính; việc này giúp doanh nghiệp có chiến lược quản trị quan hệ khách hàng, phân khúc dự án theo các phạm trù, để kiểm tra và chẩn đoán kết quả dự án hiệu quả hơn. Phân cụm là nhánh năng động trong khai phá dữ liệu.

Phân tích cụm là một chức năng khai phá dữ liệu; nó được dùng như công cụ độc lập trên phân phối dữ liệu, quan sát đặc tính của mỗi cụm, và tập trung vào tập các cụm cụ thể để phân tích tiếp. Ngoài ra, việc phân cụm được dùng như bước tiền xử lí cho thuật toán khác, như (i) lấy đặc trưng; (ii) chọn tập thuộc tính; (iii) phân lớp ... rồi thực hiện trên các cụm đã phát hiện và các thuộc tính và đặc trưng chọn được.

¹ Similar, dissimilar: tương tự, khác biệt.

² Partitioning: phân hoạch.

Do cụm là tập các đối tượng dữ liệu tương tự nhau, khác với đối tượng cụm khác, một cụm có thể xem như một lớp ngầm định. Theo nghĩa này thì việc phân cụm đôi khi được gọi là *phân lớp tự động*. Vấn đề là phân cụm tự động phát hiện các nhóm; việc này tạo nên sự khác biệt vượt trội của phân cụm.

Phân cụm cũng được gọi là *phân hoạch dữ liệu* trong vài trường hợp, do việc phân cụm đã chia bộ dữ liệu lớn thành các nhóm theo tính *tương tự*. Phân cụm có thể dùng để *phát hiện ngoại lệ*, khi ngoại lệ khác hẳn các đối tượng trong các cụm. Người ta quan tâm đến việc này, chẳng hạn phát hiện gian lận thẻ tín dụng.

Liên quan đến thống kê, việc phân cụm tập trung vào *phân tích cụm dựa trên khoảng cách*. Các công cụ phân tích cụm dựa trên k - trung bình, k - tâm đại diện¹, và vài phương pháp khác. Có các bộ phần mềm phân tích thống kê như S - Plus, SPSS và SAS. Trong học máy, phân lớp được coi là học có giám sát, liên quan đến mối quan hệ về lớp của dòng dữ liệu huấn luyện. Phân cụm được xem là *học không giám sát*, do không có thông tin về nhãn lớp. Do vậy phân cụm là dạng *học theo quan sát*, thay vì *học theo thí dụ*. Trong khai phá dữ liệu, người ta tập trung phát hiện phương pháp phân tích cụm hiệu quả trong *cơ sở dữ liệu lớn*. Do vậy, cần quan tâm đến (i) *tính mở rộng* của phương pháp phân cụm; (ii) hiệu quả khi phân cụm các *kiểu dữ liệu phức tạp*. Các kiểu dữ liệu phức tạp thể hiện qua (i) hình dạng, chẳng hạn không lồi; (ii) kiểu dữ liệu, như dữ liệu đa phương tiện; (iii) dữ liệu nhiều chiều; (iv) dữ liệu lẫn giữa dữ liệu số và dữ liệu định danh trong cơ sở dữ liệu lớn.

7.3.1.2. Các yêu cầu đối với phân tích cụm

Các yêu cầu được phân cụm ở đây là các yêu cầu đối với công cụ khai phá dữ liệu, và cũng là các khía cạnh cho phép so sánh các phương pháp phân cụm.

¹ k - mean: k - trung bình; k - medoid: k - tâm đại diện.

- *Tính mở rộng.* Nhiều thuật toán phân cụm làm việc tốt với bộ dữ liệu nhỏ, tức vài trăm dữ liệu. Với ứng dụng nhiều dữ liệu như ứng dụng Web, có hàng triệu đối tượng dữ liệu, kết quả phân cụm có thể bị lệch. Do vậy, thuật toán phân cụm cần có quy mô lớn;
- *Khả năng xử lý các dạng thuộc tính khác nhau.* Nhiều thuật toán được thiết kế để phân cụm dữ liệu số, dựa theo khoảng. Tuy nhiên các ứng dụng có các dạng dữ liệu khác nhau, như nhị phân, định danh, tức phạm trù, dữ liệu có thứ tự, dữ liệu tổng hợp. Các ứng dụng cần đến kỹ thuật phân cụm với dữ liệu phức tạp thường là xử lý ảnh, đồ thị, xử lý tư liệu;
- *Phát hiện cụm với hình dạng tùy ý.* Nhiều thuật toán phân cụm xác định cụm dựa vào độ đo Oclit hay Manhattan. Các thuật toán dựa trên độ đo khoảng cách ấy có xu hướng phát hiện cụm hình cầu, với kích thước và mật độ tương tự. Tuy nhiên cụm có thể có hình dạng khác. Chẳng hạn phân tích cụm trên dữ liệu đọc cảm biến, phát hiện ra đám cháy... thường cụm không theo dạng hình cầu. Do vậy phân cụm với hình dạng tùy ý là nhu cầu thực tế;
- *Yêu cầu về tri thức lĩnh vực để xác định tham số đầu vào.* Trong nhiều thuật toán phân cụm, người dùng cần cung cấp tri thức dưới dạng tham số đầu vào. Thuật toán nhạy cảm với tham số này, và không dễ xác định tham số này. Tri thức lĩnh vực cũng làm tăng chất lượng của phân cụm khó điều khiển;
- *Khả năng xử lý dữ liệu ồn.* Hầu hết các bộ dữ liệu thực tế đều có ngoại lai, thiếu, chưa biết, kém chính xác. Các thuật toán phân cụm có thể nhạy cảm với ồn và sẽ trả về các cụm kém chất lượng. Bởi vậy, người ta cần thuật toán phân cụm vững chắc đối với dữ liệu ồn;
- *Phân cụm gia tăng và không nhạy với thứ tự đầu vào.* Việc cập nhật sẽ tạo ra dữ liệu bổ sung đối với quá trình phân cụm. Vài thuật toán phân cụm chưa có khả năng cập nhật gia

tăng vào cấu trúc kết quả đã phân cụm, mà lại tính toán lại. Ngoài ra thuật toán có thể phụ thuộc vào thứ tự thể hiện các đối tượng. Cần thiết sử dụng thuật toán có khả năng phân cụm gia tăng và không nhạy với thứ tự của đối tượng đầu vào;

- *Khả năng phân cụm dữ liệu nhiều chiều.* Trong ứng dụng, chẳng hạn xử lý tư liệu, mỗi từ khóa ứng với một chiều dữ liệu. Thuật toán phân cụm hoạt động tốt đối với dữ liệu ít chiều, nên dữ liệu nhiều chiều là thách thức, nhất là dữ liệu thưa và lệch;
- *Phân cụm tuân theo ràng buộc.* Việc phân cụm cần tuân theo hiện trạng thể giới thực, với nhiều ràng buộc. Vấn đề ở đây là phát hiện các nhóm dữ liệu có hành vi phân cụm tốt, mà thỏa mãn các ràng buộc đặt ra;
- *Có tính sử dụng và dễ diễn giải.* Người dùng muốn kết quả phân cụm có giá trị sử dụng và phù hợp, lý giải được. Tức việc phân cụm cần gắn với thể hiện ngữ nghĩa riêng. Vậy nên cần nghiên cứu cách thích ứng ứng dụng ảnh hưởng đến việc lựa chọn các đặc trưng phân cụm và phương pháp phân cụm.

Các phương pháp phân cụm được so sánh với nhau theo các tiêu chí sau:

- *Tiêu chí phân hoạch.* Trong các phương pháp, tất cả các đối tượng được phân hoạch để trong cụm không còn phân cấp, tức là các cụm có cùng mức khái niệm. Tuy nhiên có phương pháp phân hoạch các đối tượng dữ liệu theo cách phân cấp, tức các cụm tạo nên nhiều mức ngữ nghĩa;
- *Tách biệt các cụm.* Vài phương pháp phân hoạch dữ liệu thành các cụm loại trừ nhau; nhưng cũng có phương pháp tạo nên các cụm có đối tượng chung;
- *Độ đo tương tự.* Về xác định độ tương tự, các phương pháp sử dụng (i) độ đo khoảng cách, trong không gian Oclit,

không gian vectơ, hay không gian khác; (ii) tính kết nối, dựa trên mật độ hay sự liên tục, có thể không theo khoảng cách giữa các đối tượng. Các độ đo tương tự có vai trò quan trọng trong phương pháp phân cụm. Ưu điểm của tiếp cận khoảng cách là dễ tối ưu; ưu điểm của tiếp cận mật độ hay liên tục là xử lý các hình tùy ý;

- *Không gian phân cụm.* Nếu phương pháp phân cụm quét toàn bộ không gian dữ liệu, thì chỉ phù hợp với bộ dữ liệu ít chiều. Tuy nhiên, các cụm phát hiện trong toàn bộ không gian không mấy ý nghĩa. Người ta thường xét không gian con của toàn thể. *Phân cụm trên không gian con* phát hiện các cụm và không gian con thể hiện tính tương tự của đối tượng.

Tóm lại, các thuật toán phân cụm có vài yêu cầu. Các nhân tố này gồm tính mở rộng, khả năng xử lý các dạng thuộc tính khác nhau, dữ liệu ổn, cập nhật gia tăng, xử lý hình dạng tùy ý và có ràng buộc. Các phương pháp phân cụm khác nhau theo mức phân hoạch, sự tách rời của các cụm, độ đo tương tự, sử dụng không gian con.

7.3.1.3. Tổng quan về các phương pháp phân cụm cơ bản

Có nhiều thuật toán phân cụm, nên khó tách các phương pháp phân cụm, do các phạm trù giao nhau. Tuy nhiên cũng cần phác thảo tổng quan về các phương pháp phân cụm. Nói chung, các phương pháp phân cụm căn bản được tách thành các phạm trù như sau.

Các phương pháp phân hoạch:

Định nghĩa: Phân hoạch là việc từ tập n đối tượng, kiến thiết k đoạn dữ liệu, mỗi đoạn ứng với một cụm, $k \leq n$.

Tức là chia dữ liệu thành k nhóm, mỗi nhóm có ít nhất một đối tượng. Các phương pháp phân hoạch là phân hoạch một mức trên bộ

dữ liệu. Các phương pháp điển hình là *tách biệt loại trừ nhau*. Yêu cầu này có thể nói, chẳng hạn với kĩ thuật phân hoạch mờ.

Hầu hết các phương pháp phân hoạch dựa vào khoảng cách. Với k là số đoạn được kiến thiết, phương pháp phân hoạch khởi đầu, rồi sử dụng *kĩ thuật lặp tái định vị*, nhằm cải thiện kết quả phân hoạch, nhờ việc chuyển các đối tượng từ nhóm này sang nhóm khác. Khó xác định chất lượng của việc phân hoạch. Người ta thường dùng phương pháp truyền thống với không gian con của dữ liệu; việc này có hiệu quả khi nhiều thuộc tính và dữ liệu thừa.

Người ta thường không đi đến tối ưu tổng thể đối với bài toán phân cụm theo phân hoạch, mà chấp nhận phương pháp may rủi. Các tiếp cận vét cạn, tức tham lam, như k - trung bình, k - tâm đại diện cho phép cải thiện dần dần chất lượng phân hoạch, và đạt tối ưu cục bộ. Các phương pháp may rủi làm việc tốt khi phát hiện cụm hình cầu trong cơ sở dữ liệu vừa và nhỏ. Đối với hình dạng phức tạp, cơ sở dữ liệu lớn, cần thiết mở rộng việc phân cụm dựa trên phân hoạch.

Các phương pháp phân cấp: phương pháp phân cấp tạo nên phân rã phân cấp của tập đối tượng dữ liệu. Phương pháp phân cấp có thể là (i) *mở rộng*; (ii) *chia nhỏ*, dựa trên cách phân cấp.

- Theo tiếp cận mở rộng, hay *từ dưới lên*, từ một đối tượng tạo nên nhóm tách biệt. Nó liên tục hợp nhất các đối tượng, hay nhóm, cho đến khi hợp nhất thành một nhóm. Khi đó người ta có mức trên cùng, hoặc dừng lại khi đáp ứng điều kiện dừng;
- Theo tiếp cận chia nhỏ, hay *từ trên xuống*, tất cả những đối tượng thuộc một cụm. Trong mỗi lần lặp, cụm được tách thành các cụm nhỏ hơn. Việc này dừng khi mỗi đối tượng thuộc vào một cụm, hay điều kiện dừng thỏa mãn.

Các phương pháp phân cụm phân cấp có thể dựa vào khoảng cách hay dựa vào mật độ, sự liên tục. Nhiều mở rộng của các

phương pháp sử dụng không gian con để phân cụm. Các phương pháp phân cụm không cho phép làm lại, một khi đã tách hay nhập. Nó tránh được việc tính toán quá nhiều. Các kĩ thuật này cần được cải thiện chất lượng.

Các phương pháp dựa trên mật độ: Hầu hết các phương pháp phân cụm theo phân hoạch đều dựa trên khoảng cách giữa các đối tượng. Các phương pháp như vậy chỉ phát hiện được các cụm hình cầu, và gặp khó khăn với các hình tùy ý. Có phương pháp phân cụm khác dựa vào *mật độ*, với ý tưởng rằng cụm được tạo nên nhờ số các đối tượng gần nhau đạt đến ngưỡng. Phương pháp này cho phép (i) lọc ồn, đối tượng lẻ; (ii) khai phá hình tùy ý.

Các phương pháp dựa trên mật độ có thể chia tập các đối tượng thành các cụm loại trừ nhau, tức phân cấp các cụm. Điển hình là phương pháp chỉ xét các cụm loại trừ nhau, không xét các cụm mờ. Chúng có thể làm việc với toàn không gian hay một phần không gian đối tượng.

Các phương pháp dựa vào lưới: Các phương pháp dựa vào lưới lượng hóa không gian các đối tượng thành số hữu hạn các tế bào, tạo nên cấu trúc lưới. Các thao tác phân cụm đều thực hiện trên cấu trúc lưới. Ưu điểm của tiếp cận này là thời gian xử lí nhanh, độc lập với số đối tượng dữ liệu, và chỉ phụ thuộc vào số các tế bào trong mỗi chiều trong không gian lượng hóa. Tiếp cận sử dụng lưới là hiệu quả đối với nhiều bài toán khai phá dữ liệu không gian, và phân cụm. Vậy nên, các phương pháp dựa vào lưới có thể tích hợp với các phương pháp khác, như phương pháp dựa trên mật độ hay phương pháp phân cấp.

Các phương pháp được tóm tắt lại trong bảng sau. Vài thuật toán phân cụm tích hợp ý tưởng của nhiều phương pháp phân cụm, nên khó xếp nó vào một chắc chắn; có bài toán yêu cầu tích hợp nhiều kĩ thuật phân cụm.

Bảng 7.6. Tóm tắt về phương pháp phân cụm

Phương pháp	Đặc điểm chung
Các phương pháp phân hoạch	Phát hiện các cụm loại trừ nhau, hình cầu; Dựa trên khoảng cách; Có thể dùng trung bình hay tâm đại diện để thể hiện tâm cụm; Hiệu quả với bộ dữ liệu vừa và nhỏ.
Các phương pháp phân cấp	Phân cụm với nghĩa phân cấp; Không thể chỉnh lỗi về tách, nhập; Tích hợp với kĩ thuật khác, như phân cụm vi mô, hay các liên kết đối tượng.
Các phương pháp dựa vào mật độ	Có thể phát hiện cụm với hình tùy ý; Cụm là miền cô đặc của các đối tượng không gian, tách với miền có mật độ thấp; Mật độ cụm cho biết mỗi điểm cần có số tối thiểu các điểm lân cận; Lọc được đối tượng lẻ.
Các phương pháp dựa vào lưới	Dùng cấu trúc dữ liệu lưới với nhiều độ phân giải; Thời gian xử lí nhanh.

7.3.2. Các phương pháp phân hoạch

Phiên bản phân tích cụm đơn giản và cơ bản nhất là phân hoạch, tức tổ chức các đối tượng của tập thành vài nhóm, hay cụm, tách nhau. Để xác định chính xác, giả sử đã cho trước số các cụm. Đối với các phương pháp phân hoạch, tham số này cần có trước tiên.

Cho bộ dữ liệu D , có n đối tượng, cần tạo thành k cụm. Thuật toán phân hoạch tổ chức các đối tượng thành k đoạn, $k \leq n$, trong đó mỗi đoạn thể hiện một cụm. Các cụm tạo thành để tối ưu tiêu chí phân hoạch khách quan, chẳng hạn hàm khoảng cách dựa trên sự khác nhau, sao cho các đối tượng trong một cụm sẽ tương tự nhau, và không tương tự với các đối tượng trong cụm khác; tương tự ở đây là theo các thuộc tính của bộ dữ liệu.

Sau đây là hai phương pháp phân hoạch thông dụng trung bình; (ii) k đại diện.

7.3.2.1. k trung bình

Có bộ dữ liệu D với n đối tượng trong không gian Oclit phương pháp phân hoạch phân phối các đối tượng trong D thành k cụm, $C_1, \dots, C_k$, sao cho $C_i \subset D$ và $C_i \cap C_j = \emptyset$ với $i, j = 1..k$. hàm khách quan được dùng để đánh giá chất lượng phân hoạch để các đối tượng trong cụm sẽ tương tự nhau, các đối tượng cụm sẽ khác biệt. Tức hàm khách quan dùng để nâng mức tương tự trong cụm, giảm mức tương tự giữa các cụm.

Các kĩ thuật phân hoạch dựa vào tâm sử dụng C_i như cụm. Về khái niệm, tâm cụm là điểm giữa. Người ta xác định tâm cụm theo nhiều cách, chẳng hạn trung bình hay tâm đại diện các đối tượng, được gán cho cụm. Khác nhau giữa đối tượng C_i và đại diện c_i của cụm là $d(p, c_i)$; tức khoảng cách giữa d trong không gian Oclit. Chất lượng của cụm C_i đo bằng *phương trong cụm*, tức các sai số bình phương giữa các đối tượng trong cụm và tâm c_i , $E = \sum_{i=1}^k \sum_{p \in C_i} d(p, c_i)^2$. Cả p và c_i đều được xem là nhiều chiều. Hàm khách quan này cố đạt được k cụm vừa cân bằng vừa phân tách.

Việc tối ưu sự sai khác trong cụm không đơn giản; trường hợp xấu nhất, người ta cần tính số các phân hoạch để tìm mũ của k , rồi kiểm tra các giá trị biến thiên trong cụm. Đó là bài toán NP khó trong không gian Oclit tổng quát, cho dù $k = 2$. Trong không gian Oclit hai chiều, bài toán với k cụm vẫn là NP khó. Bài toán có n đối tượng, trong không gian m chiều, với k cụm phức tạp thời gian là $O(n^{mk+1} \log n)$.

Cách thực hiện của thuật toán k trung bình (i) chọn trước k đối tượng trong D , mỗi đối tượng thể hiện một cụm; (ii) đối với các đối tượng còn lại, gán nó cho cụm nào gần nhất, khoảng cách Oclit; (iii) lặp lại để cải thiện sự biến thiên trong

đối với mỗi cụm: tính trung bình mới trên các đối tượng đã thuộc một cụm của bước lặp trước; (iv) các đối tượng được gán lại vào cụm các trung bình mới.

Thuật toán k trung bình: dùng để phân hoạch, trong đó tâm của cụm được thể hiện qua giá trị trung bình của các đối tượng trong cụm.

Đầu vào: (i) số các cụm, k ; (ii) bộ dữ liệu D với n đối tượng.

Đầu ra: tập k cụm.

Phương pháp:

Chọn tùy ý k đối tượng trong D , dùng làm tâm cụm;

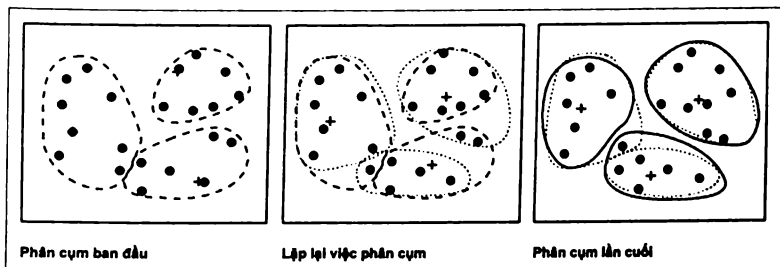
Repeat

Gán mỗi đối tượng vào cụm tương tự nhất, dựa trên giá trị trung bình của các đối tượng trong cụm;

Cập nhật trung bình cụm, tức tính giá trị trung bình của các đối tượng đối với mỗi cụm;

Until không thay đổi;

Về phân cụm nhờ phân hoạch: Xét tập các đối tượng trong không gian hai chiều, lấy $k = 3$. Theo thuật toán trên, lấy ba đối tượng làm ba tâm cụm, ghi chú trong hình là dấu "+". Mỗi đối tượng được gán vào cụm gần nhất. Tiếp theo, cập nhật tâm cụm. Lặp lại cho đến khi không thấy sự thay đổi giữa hai lần lặp.



Hình. 7.19. Phân cụm tập các đối tượng theo phương pháp k trung bình

Phương pháp k trung bình không đảm bảo hội tụ đến tối ưu toàn cục, mà chỉ dừng ở tối ưu cục bộ. Kết quả có thể phụ thuộc vào lựa chọn các đối tượng ban đầu.

Độ phức tạp về thời gian của thuật toán k trung bình là $O(nkt)$, với n đối tượng, k cụm và t lần lặp. Thông thường $k \ll n$; $t \ll n$. Do vậy phương pháp này là hiệu quả được bộ dữ liệu lớn.

Có vài phiên bản của phương pháp k trung bình. Chúng khác nhau về (i) cách chọn k đối tượng ban đầu; (ii) tính độ khác nhau; (iii) chiến lược tính trung bình cụm.

Phương pháp k trung bình có thể áp dụng chỉ khi trung bình của tập các đối tượng là xác định. Trường hợp không áp dụng được là khi ứng dụng có dữ liệu mang thuộc tính định danh. Phương pháp k -kiểu¹ là biến thể của k trung bình, để dùng cho dữ liệu định danh; người ta thay trung bình cụm bằng các kiểu. Nó sử dụng (i) độ đo sự khác biệt mới, để xử lý các đối tượng định danh; (ii) phương pháp dựa trên tần suất, để cập nhật các kiểu của cụm. Các phương pháp k trung bình và k -kiểu có thể kết hợp để xử lý dữ liệu cả số lẫn định danh.

Việc nhất thiết xác định trước số k là nhược điểm của phương pháp. Người ta dùng các cách để xác định k nào là tốt đối với bài toán phân hoạch. Phương pháp k trung bình không phù hợp đối với việc khám phá các cụm với hình không lối, hay các cụm với hình dạng khác nhau. Ngoài ra, nó nhạy cảm với ồn và dữ liệu lẻ, do một vài dữ liệu như thế sẽ ảnh hưởng đến giá trị trung bình.

Để mở rộng thuật toán k trung bình, có tiếp cận (i) sử dụng tập kích thước tốt của các mẫu trong phân cụm; (ii) tiếp cận khác sử dụng chỉ mục phân cấp không gian để ghi lại chi phí khi tính trung bình; (iii) dùng ý tưởng cụm vi mô, nhóm các đối tượng vào cụm vi mô, rồi thực hiện phân cụm k trung bình trong cụm vi mô đó.

¹ k - mode: k -kiểu.

7.3.2.2. *k* tâm đại diện

Thuật toán *k* trung bình nhạy cảm với dữ liệu lẻ, do các đối tượng này xa số đông các đối tượng khác, nên khi đưa nó vào cụm, giá trị trung bình sẽ biến dạng.

Về nhược điểm của k trung bình: Xét sáu điểm không gian một chiều, có giá trị 1, 2, 3, 8, 9, 10 và 25. Dễ thấy có hai cụm {1, 2, 3} và {8, 9, 10}; điểm 25 là bên ngoài, xem như lẻ. Tuy nhiên thuật toán *k* trung bình với $k = 2$, thì phân hoạch {{1, 2, 3}, {8, 9, 10, 25}} có phương sai trong cụm là $(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 + (8 - 13)^2 + (9 - 13)^2 + (10 - 13)^2 + (25 - 13)^2 = 196$, với hai trung bình là 2 và 13. So sánh với phân hoạch {{1, 2, 3, 8}, {9, 10, 25}}, có hai trung bình là 3.5 và 14.67, phương sai là $(1 - 3.5)^2 + \dots + (25 - 14.67)^2 = 189.67$. Tâm của cụm thứ hai là 14.67, cách xa các phần tử trong cụm. Tuy nhiên, phân hoạch đầu là hợp lí hơn.

Liên quan đến việc thay đổi thuật toán *k* trung bình để giảm mức nhạy cảm với dữ liệu lẻ, thay vì lấy giá trị trung bình của các đối tượng trong cụm, người ta xem các đối tượng trong cụm thể hiện cụm, dùng một đối tượng đại diện cho cụm. Các đối tượng được gán vào cụm có đối tượng thể hiện tương tự nhất. Phương pháp phân hoạch thực hiện dựa trên nguyên tắc tối thiểu sai khác giữa mỗi đối tượng *p* và đối tượng đại diện. Tức người ta sử dụng

tiêu chí sai tuyệt đối, xác định theo $E = \sum_{i=1}^k \sum_{p \in C_i} d(p, o_i)^2$. Đây là cơ sở

của phương pháp *k* tâm đại diện. Khi $k = 1$, người ta có thể phát hiện được trung điểm với độ phức tạp thời gian $O(n^2)$. Với *k* khác, bài toán *k* tâm đại diện là NP khó.

Thuật toán phân hoạch quanh tâm đại diện PAM1 là dạng thông dụng của phân cụm *k* tâm đại diện. Nó giải quyết vấn đề theo cách lặp, tham lam. Như thuật toán *k* trung bình, thuật toán chọn các đối tượng đại diện ngẫu nhiên. Tất cả các phương án đại

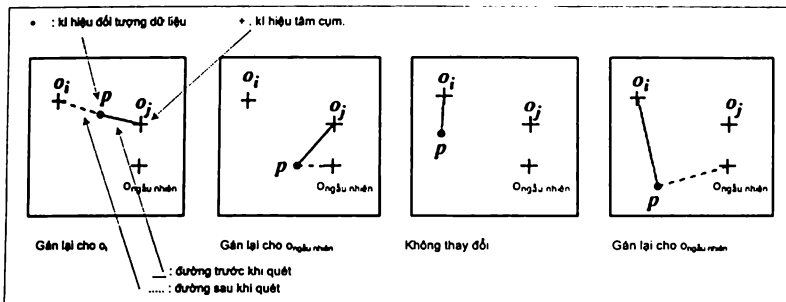
¹ PAM: partitioning around medoids: phân hoạch quanh tâm đại diện.



diện được thử, để biết đối tượng nào làm đại diện là tốt, cho đến khi không cải thiện chất lượng được nữa.

Giả sử đã chọn các tâm đại diện là $o_1, o_2, \dots, o_k$. Để xác định đối tượng o có thể thay thế o_i để tăng chất lượng phân cụm, cần xét các đối tượng p đối với các tâm mới $\{o_1, o_2, \dots, o_{i-1}, o, o_{i+1}, \dots, o_k\}$. Có các khả năng đối với đối tượng p :

- Trước đây p thuộc cụm o_i , nếu p gần với o_i , nó sẽ được chuyển về cụm o_i . Ngược lại p thuộc vào cụm có tâm đại diện mới, o ;
- Trước đây p thuộc cụm o_i , nếu p gần với tâm mới o , nó sẽ chuyển về cụm o . Ngược lại, p vẫn thuộc cụm o_i .



Hình 7.19. Các trường hợp đối tượng p thuộc về các cụm mới

Mỗi lần thay đổi tâm đại diện, sai số tuyệt đối E được tính lại; nó được dùng cho hàm giá. Hàm giá sẽ tính sai khác về giá trị sai số tuyệt đối giữa lần sau và lần trước. Nếu giá trị hàm này âm, tức E giảm, người ta chấp nhận thay o_i bằng o .

So sánh hai phương pháp k trung bình và k tâm đại diện, người ta thấy phương pháp k tâm đại diện là mạnh hơn đối với dữ liệu ổn, dữ liệu lể, do nó ít bị ảnh hưởng của dữ liệu lể hay các giá trị quá xa. Tuy nhiên, độ phức tạp của mỗi lần lặp trong thuật toán k tâm đại diện là $O(k(n - k)^2)$. Với giá trị n, k lớn, lượng tính toán lớn, tốn kém so với phương pháp k trung bình. Cả hai phương pháp đều yêu cầu người dùng xác định số các cụm, k .

Về việc mở rộng phương pháp k tâm đại diện, người ta thấy thuật toán phân hoạch k tâm đại diện như PAM là hiệu quả đối với bộ dữ liệu nhỏ, nhưng không tốt đối với bộ dữ liệu lớn. Đối với bộ dữ liệu lớn, có phương pháp dựa trên mẫu CLARA¹. Thay vì xét toàn bộ dữ liệu, phương pháp này sử dụng mẫu dữ liệu ngẫu nhiên. Thuật toán PAM được dùng để tính tâm đại diện tốt nhất từ mẫu này. Thật lí tưởng nếu mẫu này thể hiện sát với bộ dữ liệu gốc. Trong nhiều trường hợp, một mẫu lớn hoạt động tốt nếu được tạo sao cho mỗi đối tượng có xác suất chọn vào mẫu là đều nhau. Các đối tượng đại diện được chọn sẽ tương tự như các đối tượng tạo nên từ toàn thể dữ liệu. CLARA thực hiện phân cụm với nhiều bộ mẫu ngẫu nhiên, và trả về giải pháp phân cụm tốt nhất. Độ phức tạp để tính các tâm đại diện trên mẫu ngẫu nhiên là $O(ks^2 + k(n - k))$, trong đó s là kích thước mẫu. CLARA cho phép xử lí bộ dữ liệu lớn hơn, so với PAM.

Thuật toán PAM, cho phép phân hoạch dựa trên đối tượng tâm đại diện hay trung tâm.

Đầu vào: (i) số các cụm, k ; (ii) bộ dữ liệu D với n đối tượng.

Đầu ra: tập k cụm.

Phương pháp:

Chọn tùy ý k đối tượng trong D làm các đối tượng đại diện khởi đầu;

Repeat

 Gán mỗi đối tượng còn lại vào cụm có đối tượng đại diện gần nhất;

 Chọn ngẫu nhiên đối tượng o_j ;

 Tính tổng giá, S , khi thay đối tượng o_j bằng o ;

 Nếu $S < 0$ thì thay o_j bằng o , tạo nên tập k tâm đại diện mới;

Until không thay đổi.

¹ CLARA Clustering LARge Application: ứng dụng phân cụm dữ liệu lớn.

Hiệu quả của CLARA tùy vào kích thước mẫu. PAM tìm các tâm đại diện tốt nhất trong bộ dữ liệu đã biết, trong khi CLARA tìm tâm đại diện tốt nhất trong *mẫu đã chọn* của bộ dữ liệu. CLARA không thể tìm thấy phân cụm tốt, nếu các tâm đại diện lấy mẫu tốt nhất nằm xa tâm đại diện tốt nhất thực sự.

Về việc cải thiện thuật toán CLARA, người ta thấy khi tìm kiếm các tâm đại diện tốt nhất, PAM kiểm tra mọi đối tượng trong bộ dữ liệu với mọi tâm hiện tại, nhưng CLARA chỉ xét riêng các tâm ứng viên với chỉ mẫu ngẫu nhiên của bộ dữ liệu. Thuật toán CLARANS¹ cho thấy ích lợi giữa giá và hiệu quả khi sử dụng các mẫu trong phân cụm. (i) Trước tiên nó chọn ngẫu nhiên k đối tượng trong bộ dữ liệu làm tâm đại diện. (ii) Chọn ngẫu nhiên tâm hiện tại x và đối tượng y ; y không là tâm đại diện hiện tại. (iii) Kiểm tra khả năng thay x bằng y ; nếu tốt hơn, thực hiện thay thế. Nó thực hiện t lần tìm kiếm ngẫu nhiên. Tập các tâm hiện tại sau t bước được xem là cực đại địa phương. Thuật toán lặp quá trình này m lần, lấy kết quả cuối như giải pháp phân cụm tối ưu.

7.3.3. Các phương pháp phân cấp

Các phương pháp phân hoạch đáp ứng các yêu cầu phân cụm về tổ chức tập các đối tượng thành một số nhóm rời nhau, nhưng trong vài trường hợp, cần thiết phân hoạch dữ liệu thành các nhóm theo các mức khác nhau, tức phân cấp. Biểu diễn các đối tượng dữ liệu theo dạng phân cấp là có ích đối với tóm tắt và hiển thị dữ liệu.

Định nghĩa: Phương pháp phân cụm phân cấp² là phương pháp cho phép nhóm các đối tượng dữ liệu thành các cấp độ, hay cây của các cụm.

¹ CLARANS: Clustering Large Application based on RANdomized Search: CLARA: dựa trên tìm kiếm ngẫu nhiên.

² Hierarchical clustering method: phương pháp phân cụm phân cấp.

Dù người ta phân hoạch dữ liệu một cách phân cấp, không cần giả thiết các dữ liệu đó theo cấu trúc phân cấp. Việc sử dụng phân cấp chỉ để tóm tắt và thể hiện dữ liệu theo một cách gọn hơn.

Về phương pháp phân cụm phân cấp: (i) *phân cụm gộp* cho phép tổ chức thành phân cấp theo chiến lược dưới lên; (ii) *phân cụm tách* cho phép tổ chức theo chiến lược trên xuống. Các phương pháp gộp bắt đầu với các đối tượng đơn, rồi nhập lại để được cụm lớn hơn. Các phương pháp tách bắt đầu với toàn thể dữ liệu, rồi tách nó thành các cụm nhỏ hơn.

Các phương pháp phân cụm phân cấp sẽ gặp khó khăn trong việc lựa chọn giữa gộp hay tách. Một khi đã theo một tiếp cận, người ta không được phép quay lại. Quyết định này cũng ảnh hưởng đến chất lượng phân cụm. Các phương pháp này không dùng cho nhiều dữ liệu được, do quyết định gộp hay tách cần kiểm tra và đánh giá nhiều đối tượng, nhiều cụm.

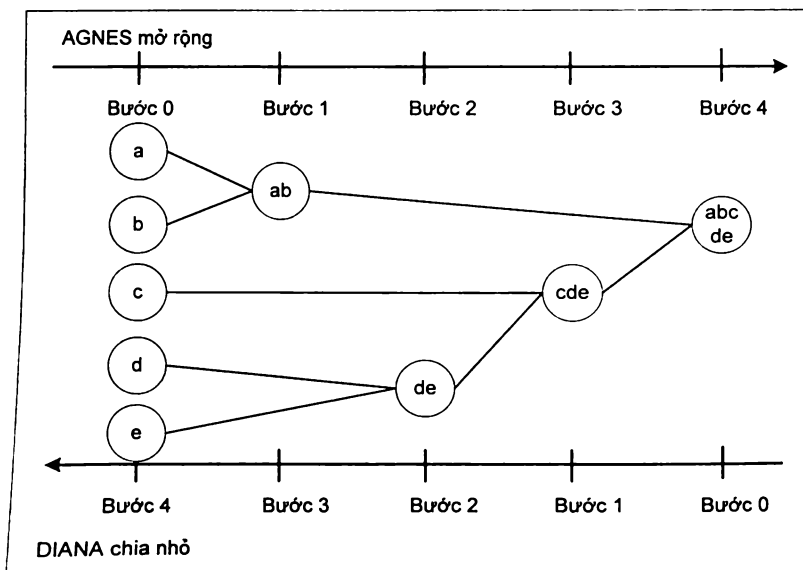
Người ta có thể kết hợp nhiều phương pháp phân cụm khi thực hiện phân cụm phân cấp, tạo thành *phân cụm nhiều pha*. Có nhiều cách nhìn nhận về các phương pháp phân cụm phân cấp, có thể là:

1. *Phương pháp thuật toán*. Các phương pháp phân cụm gộp, tách, hay nhiều pha thuộc vào các phương pháp thuật toán, với nghĩa chúng xét các đối tượng dữ liệu như những thứ tất định, và tính các cụm theo khoảng cách tất định giữa các đối tượng;
2. *Phương pháp xác suất*. Người ta sử dụng mô hình xác suất để tạo các cụm và đo chất lượng của cụm theo mức phù hợp của mô hình;
3. *Phương pháp Bayes*. Phương pháp này tính phân phối của các khả năng phân cụm. Tức là thay vì đưa ra việc phân cụm tất định đơn trên bộ dữ liệu, chúng cho nhóm các cấu trúc phân cụm và các xác suất, có điều kiện theo bộ dữ liệu đã biết.

7.3.3.1. So sánh phân cụm phân cấp gộp với phân cụm phân cấp tách

Phương pháp phân cụm phân cấp có thể (i) gộp; hay (ii) tách, tùy theo việc phân rã phân cấp được tạo nên theo cách (i) *dưới lên*, tức gộp; hay (ii) *trên xuống*, tức tách.

- *Phương pháp phân cụm phân cấp gộp* dùng chiến lược gộp dưới lên. Nó bắt đầu từ việc cho mỗi được tạo thành một cụm, rồi nhiều lần gộp các cụm nhỏ để được cụm lớn hơn. Quá trình dừng khi (i) tất cả các đối tượng thuộc về cụm đơn; hay (ii) thỏa mãn điều kiện dừng. Cụm đơn trở thành gốc của cây phân cấp. Đối với mỗi lần gộp, người ta kết hợp hai cụm gần nhau. Như vậy, quá trình này cần n bước lặp;
- *Phương pháp phân cụm phân cấp tách* sử dụng chiến lược trên xuống. Ban đầu tx các đối tượng thuộc về một cụm; đó là gốc cây phân cấp. Rồi phương pháp chia cụm gốc thành vài cụm nhỏ hơn, cho đến khi mỗi cụm chỉ là một đối tượng. Quá trình có thể dừng khi mỗi cụm đủ kết hợp nhau.



Hình 7.20. Phân cụm phân cấp gộp và tách trên bộ dữ liệu {a, b, c, d, e}

Trong hai tiếp cận trên, người dùng có thể xác định số các cụm cần thiết để kết thúc và lấy làm kết quả phân cụm.

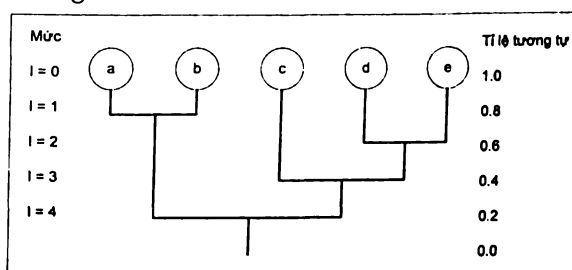
Vẽ so sánh phân cụm gộp với tách: hình cho thấy ứng dụng của phương pháp phân cụm phân cấp gộp AGNES¹ và phương pháp phân cụm phân cấp tách DIANA² đối với bộ dữ liệu $\{a, b, c, d, e\}$.

Khởi đầu, AGNES đặt mỗi đối tượng vào một cụm, rồi kết hợp các cụm. Quá trình kết thúc sau năm bước. Khoảng cách Oclit giữa hai đối tượng của hai cụm khác nhau sẽ xác định việc sát nhập hai cụm. Đây là tiếp cận *liên kết đơn*, cho biết mỗi cụm được thể hiện qua nhiều đối tượng, độ tương tự giữa hai cụm được đo bằng độ tương tự của cặp *gần nhất* của các đối tượng dữ liệu.

Phương pháp DIANA xử lí theo cách ngược lại. Ban đầu có một cụm, rồi nó được tách theo nguyên tắc, chẳng hạn khoảng cách Oclit cực đại giữa hai dữ liệu gần nhau nhất trong cụm này. Quá trình tách lặp cho đến khi cụm mới có chỉ một đối tượng.

Định nghĩa: *Lược đồ cây*³ là cây cho phép thể hiện quá trình phân cụm phân cấp.

Lược đồ cây là thông dụng, cho biết cách nhóm các đối tượng đối với phân cụm gộp, hay cách phân hoạch đối với phân cụm tách, theo từng bước.



Hình 7.21. Lược đồ cây thể hiện phân cụm phân cấp của bộ dữ liệu $\{a, b, c, d, e\}$

¹ AGNES: AGglomerative NESTing: lồng gộp lại.

² DIANA: Divisive ANALysis: phân tích tách.

³ Dendrogram: lược đồ cây.

7.3.3.2. Độ đo khoảng cách trong các phương pháp thuật toán

Khi sử dụng phương pháp gộp hay tách, người ta cần đo khoảng cách giữa hai cụm; mỗi cụm là tập các đối tượng. Bốn độ đo dùng rộng rãi. Ký hiệu p, p' là hai đối tượng, hay điểm; $|p - p'|$ là khoảng cách giữa p, p' ; m_i là tâm cụm C_i ; n_i là số các đối tượng trong C_i .

- Khoảng cách tối thiểu: $d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$;
- Khoảng cách tối đa: $d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$;
- Khoảng cách tâm cụm: $d_{\text{tâm}}(C_i, C_j) = |m_i - m_j|$;
- Khoảng cách trung bình: $d_{\text{tb}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$;

Khi thuật toán sử dụng khoảng cách trung bình $d_{\text{tb}}(C_i, C_j)$ để đo khoảng cách giữa các cụm, thuật toán này được gọi là *thuật toán phân cụm láng giềng gần nhất*. Nếu quá trình phân cụm dừng với điều kiện các cụm gần nhất có khoảng cách vượt ngưỡng xác định, thuật toán này được gọi là *thuật toán liên kết đơn*¹.

Nếu xem cụm như nút của cây, với các cạnh là khoảng cách giữa hai cụm, thì việc gộp hai cụm cũng như việc nhập hai nút cây. Vậy thuật toán phân cụm phân cấp gộp sử dụng độ đo khoảng cách tối thiểu được gọi là *thuật toán cây bao trùm tối thiểu*²; trong đó cây bao trùm là cây nối tất cả các đỉnh, và cây bao trùm tối thiểu là cây có tổng các trọng số trên cạnh là nhỏ nhất.

Khi sử dụng khoảng cách tối đa $d_{\max}(C_i, C_j)$, thuật toán được gọi là *thuật toán phân cụm láng giềng xa nhất*. Nếu việc phân cụm dừng khi khoảng cách cực đại giữa các cụm vượt quá ngưỡng xác định, nó được gọi là *thuật toán liên kết đầy đủ*³. Với nhìn nhận các cụm như nút trên cây, người ta thấy mỗi cụm như đồ thị con đầy

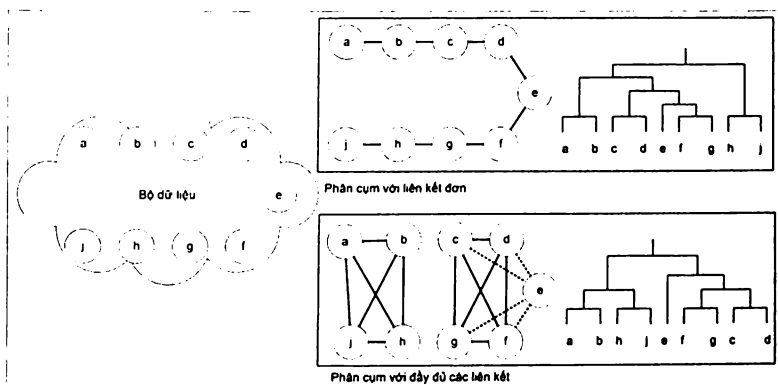
¹ Single - linkage algorithm: thuật toán liên kết đơn.

² Minimal spanning tree algorithm: thuật toán cây bao trùm tối thiểu.

³ Complete - linkage algorithm: thuật toán liên kết đầy đủ.

đủ; các cạnh nối các nút trong các cụm. Khoảng cách giữa hai cụm được xác định theo các nút cách xa nhất trong hai cụm. Các thuật toán láng giềng xa nhất có xu hướng tối thiểu hóa việc tăng đường kính của các cụm trong mỗi lần lặp. Nếu các cụm thực tế có kích thước xấp xỉ nhau và khá nén chặt, phương pháp này sẽ cho các cụm chất lượng cao. Ngược lại, các cụm sẽ không mấy ý nghĩa.

Các độ đo cực đại và cực tiểu thể hiện hai thái cực về đo khoảng cách giữa các cụm. Chúng có xu hướng quá nhạy với dữ liệu lẻ hay ồn. Việc sử dụng khoảng cách trung bình hay khoảng cách tâm là giải pháp dung hòa, khắc phục được trường hợp có dữ liệu lẻ hay ồn. Tiếp cận với khoảng cách tâm sẽ đơn giản về tính toán. Tiếp cận với khoảng cách trung bình có ưu điểm về xử lý dữ liệu phạm trù hay dữ liệu số.



Hình 7.22. Phân cụm phân cấp theo các liên kết đơn và đầy đủ

Về thuật toán liên kết đơn so với liên kết đầy đủ: Xét bộ dữ liệu được thể hiện theo (i) phân cụm phân cấp đơn, với lược đồ cây và cây liên kết đơn; để đơn giản, các cung giữa các cụm $\{A, B, J, H\}$ và $\{C, D, G, E, F\}$ đã bỏ đi; (ii) phân cụm liên kết đầy đủ. Thí dụ cho thấy việc dùng liên kết đơn cho phép tìm thấy các cụm được xác định theo sự gần cục bộ; liên kết đầy đủ cho phép tìm thấy các cụm thiên về việc đóng tổng thể.

7.3.3.3. Phân cụm phân cấp nhiều pha, dùng cây đặc trưng phân cụm

BIRCH¹ được thiết kế để phân cụm lượng lớn dữ liệu số, nhờ tích hợp phân cụm phân cấp, ở pha *phân cụm nhỏ*², và các phương pháp phân cụm khác, như phân hoạch lặp, ở pha *phân cụm lớn*. Nó khắc phục hai vấn đề trong phương pháp phân cụm gộp về (i) quy mô; (ii) khả năng quay trở lại.

Nó dùng khái niệm về *đặc trưng phân cụm* để tóm tắt một cụm, và *cây đặc trưng phân cụm*, kí hiệu cây CF, để biểu diễn một phân cấp cụm. Các cấu trúc này giúp phương pháp phân cụm đạt được tốc độ và sự mở rộng tốt, trong cơ sở dữ liệu lớn hay cơ sở dữ liệu gai tăng, và cũng hiệu quả đối với phân cụm gia tăng hay phân cụm động trên các đối tượng đầu vào.

Xét cụm gồm các đối tượng dữ liệu n chiều. Đặc trưng phân cụm CF của cụm này là vectơ 3 chiều, cho thông tin về các cụm đối tượng. $CF = (n, LS, SS)$; trong đó LS là tổng tuyến tính của n điểm, tức $LS = \sum_{i=1}^n x_i$, và SS là tổng bậc hai, tức $SS = \sum_{i=1}^n x_i^2$.

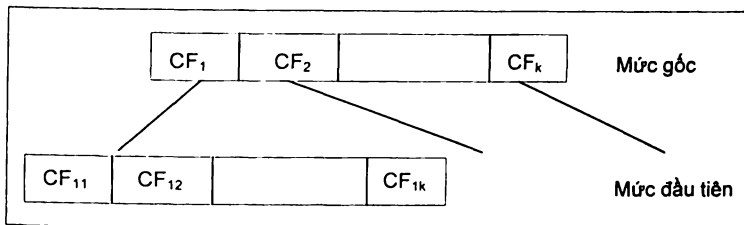
Đặc trưng phân cụm là tóm lược về thống kê đối với cụm đã biết. Sử dụng nó, người ta có thể rút ra các thống kê khác. Việc tóm lược một cụm nhờ đặc trưng phân cụm sẽ tránh được việc lưu thông tin về các đối tượng, hay điểm, riêng. Đó chính là ưu điểm của BIRCH trong không gian. Ngoài ra, các đặc trưng phân cụm có tính *cộng vào*. Tức đối với hai cụm C_1, C_2 , với các đặc trưng phân cụm CF_1, CF_2 , thì khi hợp nhất hai cụm này, $CF = CF_1 + CF_2 = (n_1 + n_2, LS_1 + LS_2, SS_1 + SS_2)$.

Về đặc trưng phân cụm: giả sử có ba điểm (2, 5), (3, 2), và (4, 2) trong cụm C_1 . Đặc trưng phân cụm của C_1 là $CF_1 = (3, (2 + 3 + 4, 5 + 2 + 3), (2^2 + 3^2 + 4^2, 5^2 + 2^2 + 3^2)) = (3, (9, 10), (29, 38))$. Giả sử C_1 tách biệt cụm thứ hai, C_2 . C_2 có $CF_2 = (3, (35, 36), (417, 440))$. Khi hợp

¹ BIRCH: Balanced Iterative Reducing and Clustering using Hierarchie: giảm lặp cân đối và phân cụm theo phân cấp.

² Microclustering, macroclustering: phân cụm nhỏ, phân cụm lớn.

nhất C_1 và C_2 , được C_3 , người ta được $CF_3 = (3 + 3, (9 + 35, 10 + 36), (29 + 417, 38 + 440)) = (6, (9, 10), (446, 478))$.



Hình 7.23. Cấu trúc cây CF

Định nghĩa: Cây CF là cây cân đối theo chiều cao, chứa các đặc trưng phân cụm đối với phân cụm phân cấp.

Trên cây CF, nút không lá có các nút con. Nút không lá chứa các tổng của CF của các nút con. Cây CF có hai tham số (i) *nhân tố chia nhánh B*; (ii) *ngưỡng T*. Nhân tố chia nhánh xác định số cực đại các nút con của nút không lá. Tham số ngưỡng xác định đường kính cực đại của các cụm con lưu tại các nút lá trên cây. Hai tham số này ngầm điều khiển kích thước cây.

Biết lượng hạn chế của bộ nhớ chính, điều quan trọng trong BIRCH là tối thiểu thời gian vào/ra. Nó áp dụng kỹ thuật phân cụm *nhiều pha*. Bộ dữ liệu được quét để thu được phân cụm tốt, và các lần quét sau sẽ cải thiện chất lượng phân cụm. Các pha là:

- *Pha 1.* BIRCH quét cơ sở dữ liệu để xây dựng cây CF ban đầu, trong bộ nhớ. Việc này xem như nén dữ liệu nhiều mức, mà vẫn giữ cấu trúc phân cụm vốn có;
- *Pha 2.* Áp dụng thuật toán phân cụm, đã chọn, để phân cụm các nút lá của cây CF. Nó loại bỏ các cụm thừa, như các dữ liệu lẻ, và nhóm các cụm đậm đặc vào cụm lớn hơn.

Trong pha 1, cây CF được kiến thiết động, khi bổ sung các đối tượng. Phương pháp theo tiếp cận gia tăng. Một dữ liệu được bổ sung vào mục lá cây gần nhất, tức cụm con. Nếu đường kính của cụm lớn hơn giá trị ngưỡng, nút lá này và có thể cả nút khác được

tách. Sau khi bổ sung đối tượng mới, thông tin về đối tượng này được chuyển lên gốc cây. Kích thước cây CF có thể thay đổi do thay đổi ngưỡng. Nếu kích thước bộ nhớ cần lưu trữ cây CF vượt kích thước được phép trong bộ nhớ, cần sử dụng ngưỡng lớn hơn, và cây CF được kiến thiết lại.

Quá trình xây dựng lại được thực hiện nhờ kiến thiết cây mới, từ các nút lá của cây cũ. Vậy, quá trình này không cần xem lại tất cả các đối tượng, hay điểm. Điều này tương tự như bổ sung hay tách khi kiến thiết cây B+. Do vậy, để xây dựng cây này, chỉ cần một lần đọc dữ liệu. Người ta sử dụng nhiều phương pháp và may rủi để xử lý dữ liệu lể và nâng chất lượng cây CF, nhờ quét thêm các dữ liệu. Một khi cây CF được xây dựng, bất kì thuật toán phân cụm, chẳng hạn thuật toán phân hoạch điển hình, đều có thể dùng cho cây CF tại pha 2.

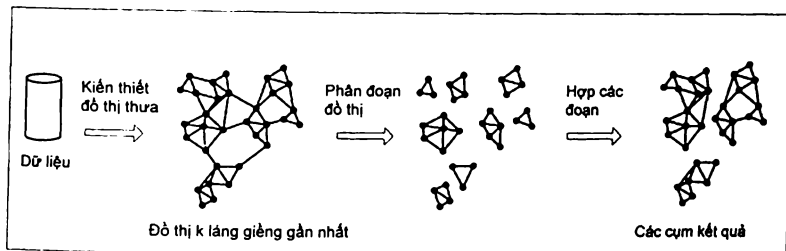
Về tính hiệu quả của BIRCH, người ta thấy độ phức tạp thời gian của thuật toán là $O(n)$, trong đó n là số đối tượng phân cụm. Kinh nghiệm cho thấy khả năng mở rộng tuyến tính của thuật toán theo số các đối tượng. Tuy nhiên, do mỗi nút trong cây CF chỉ có thể giữ số có hạn các mục, theo kích thước của nó, một nút cây CF không luôn ứng với điều người dùng nghĩ về cụm thực tế. Ngoài ra, nếu các cụm không luôn có dạng cầu, thì BIRCH không chạy tốt, do nó sử dụng khái niệm đường kính để điều khiển biên cụm.

Ý tưởng sử dụng đặc trưng phân cụm và cây CF có ý nghĩa trong việc phân cụm với dữ liệu gia tăng, động.

7.3.3.4. Phân cụm phân cấp nhiều pha, dùng mô hình hóa động

Năm 1999 người ta đề xuất thuật toán Chameleon, để phân cụm phân cấp, dùng mô hình hóa động để xác định độ tương tự giữa cặp cụm. Trong thuật toán, độ tương tự cụm được đánh giá theo (i) cách các đối tượng trong cụm liên kết nhau; (ii) độ gần của các cụm. Tức là hai cụm được hợp lại nếu *tính liên kết* giữa chúng là cao, và chúng *gần nhau*. Vậy, thuật toán không phụ thuộc vào

mô hình tĩnh, do người dùng cung cấp, và có thể tự thích nghi với tính chất bên trong của các cụm được hợp nhất. Quá trình hợp nhất là thuận tiện đối với việc phát hiện các cụm đồng nhất tự nhiên và áp dụng cho tất cả các kiểu dữ liệu, do xác định được hàm tương tự.



Hình 7.24. Thuật toán Chameleon phân cụm phân cấp, dựa trên mô hình hóa động

Trong hình cho thấy thuật toán Chameleon dùng tiếp cận đồ thị k láng giềng gần nhất để kiến thiết đồ thị thưa, trong đó mỗi đỉnh ứng với một đối tượng, cạnh nối hai đỉnh, tức đối tượng, nếu một đối tượng thuộc vào các đối tượng k tương tự nhất đối tượng này. Các cạnh có trọng số, phản ánh độ tương tự giữa các đối tượng. Chameleon sử dụng thuật toán phân hoạch để tách đồ thị thành các cụm con, và đã giảm số cạnh. Cụm C được tách thành C_i và C_j sao cho tối thiểu *trọng số của các cạnh cần được loại bỏ*. Nó đánh giá tính liên kết *tuyệt đối* giữa cụm C_i và C_j .

Rồi Chameleon dùng thuật toán phân cụm phân cấp mở rộng, cho phép lặp lại việc hợp nhất các cụm con, dựa vào độ tương tự của chúng. Để xác định cặp các cụm con tương tự nhất, nó tính đến cả tính liên kết và tính gần của các cụm. Đặc biệt, Chameleon xác định độ tương tự giữa cặp cụm C_i và C_j theo *tính liên kết tương đối* $RI(C_i, C_j)$ và *tính gần tương đối* $RC(C_i, C_j)$.

Định nghĩa: Tính liên kết tương đối giữa hai cụm C_i, C_j là liên kết tuyệt đối giữa chúng, được chuẩn hóa theo liên kết bên trong của hai cụm.

$$RI(C_i, C_j) = \frac{|EC_{\{C_i, C_j\}}|}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)},$$

trong đó $EC\{C_i, C_j\}$ là cạnh cắt đã xác định đối với cụm chứa cả C_i và C_j . Tương tự ECC_i là tổng tối thiểu của các cạnh cắt mà phân hoạch C_i thành hai phần bằng nhau.

Định nghĩa: Độ gần tương đối giữa cặp cụm C_i và C_j là độ gần tuyệt đối được chuẩn hóa theo độ gần bên trong hai cụm.

$$RC(C_i, C_j) = \frac{\bar{S}_{EC\{C_i, C_j\}}}{\frac{|C_i|}{|C_i| + |C_j|} \bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i| + |C_j|} \bar{S}_{EC_{C_j}}},$$

trong đó $\bar{S}_{EC\{C_i, C_j\}}$ là trọng số trung bình của các cạnh nối các đỉnh C_i đến các đỉnh trong C_j , và $\bar{S}_{EC_{C_i}}$ là trọng số trung bình của các cạnh thuộc về phần bị cắt của cụm C_i .

7.3.3.5. Phân cụm phân cấp xác suất

Các phương pháp phân cụm phân cấp theo thuật toán sử dụng các độ đo liên kết là dễ hiểu và thường phân cụm hiệu quả. Tuy nhiên, các phương pháp đó có thể gặp vài trở ngại. (i) Việc chọn độ đo khoảng cách tốt để phân cụm phân cấp chưa xuất sắc. (ii) Áp dụng phương pháp thuật toán không chấp nhận các đối tượng dữ liệu thiếu các giá trị thuộc tính. Khi dữ liệu được quan sát một phần, tức thiếu vài giá trị thuộc tính, không dễ phân cụm phân cấp theo thuật toán, do không tính được khoảng cách. (iii) Hầu hết các phương pháp phân cụm phân cấp là theo kinh nghiệm, tìm các quyết định tách, nhập tại mỗi bước. Vậy nên đích tối ưu của phân cấp cụm không rõ ràng.

Phân cụm phân cấp xác suất nhằm khắc phục các nhược điểm trên, nhờ dùng các mô hình xác suất để tính khoảng cách giữa các cụm.

Một cách nhìn nhận bài toán phân cụm là xét tập các đối tượng dữ liệu được phân cụm như mẫu của cơ chế tạo dữ liệu,

hay như *mô hình sinh*. Chẳng hạn khi đưa việc phân tích cụm trên tập điều tra tiếp thị, người ta giả sử các điều tra thu thập là một mẫu về ý kiến của tất cả các khách hàng. Ở đây, cơ chế tạo sinh dữ liệu là phân phối xác suất các ý kiến, ứng với khách hàng khác nhau, mà không thể thu được trực tiếp, đầy đủ. Nhiệm vụ phân cụm được ước lượng mô hình sinh chính xác có thể được, theo các đối tượng dữ liệu quan sát cần được phân cụm.

Thực tế, người ta có thể giả thiết rằng các mô hình sinh dữ liệu chấp nhận các hàm phân phối thông thường, như phân phối Gauss hay phân phối Bernoulli, có tham số điều khiển. Nhiệm vụ học mô hình sinh đưa về việc phát hiện các giá trị tham số, để mô hình phù hợp nhất với bộ dữ liệu quan sát.

Về *mô hình sinh*: giả sử có các điểm một chiều $X = \{x_1, \dots, x_n\}$ để phân tích cụm. Giả sử các điểm dữ liệu được tạo theo phân phối

Gauss $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, trong đó có các tham số trung

bình μ và phương sai σ^2 . Xác suất để điểm $x_i \in X$ được mô hình tạo

là $P(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$. Do vậy, khả năng X được mô

hình sinh ra là $L(N(\mu, \sigma^2) : X) = P(X | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$.

Nhiệm vụ học mô hình sinh là phát hiện các tham số μ và σ^2 sao cho khả năng $L(N(\mu, \sigma^2) : X)$ cực đại, tức phát hiện $N(\mu_0, \sigma_0^2) = \arg \max\{L(N(\mu, \sigma^2) : X)\}$, trong đó $\max\{L(N(\mu, \sigma^2) : X)\}$ gọi là *khả năng cực đại*.

Cho tập các đối tượng, chất lượng của cụm tạo nên nhờ tất cả các đối tượng có thể được đo bằng khả năng cực đại này. Đối với tập các đối tượng được phân hoạch thành m cụm $C_1, \dots, C_m$, chất lượng có thể được đo bằng $Q(\{C_1, \dots, C_m\}) = \prod_{i=1}^m P(C_i)$, trong đó $P()$ là khả năng cực đại. Nếu hợp hai cụm C_i và C_j được cụm $C_i \cup C_j$, thì chất lượng thay đổi của phân cụm chung là

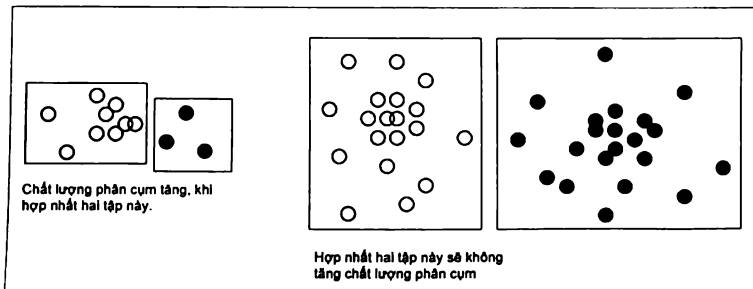
$$\begin{aligned}
& Q(\{(C_1, \dots, C_m) - \{C_i, C_j\}\} \cup \{C_i \cup C_j\}) - Q(\{(C_1, \dots, C_m)\}) \\
&= \frac{\prod_{k=1}^m P(C_k) \cdot P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)} - \prod_{k=1}^m P(C_k) \\
&= \prod_{k=1}^m P(C_k) \cdot \left(\frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)} - 1 \right).
\end{aligned}$$

Khi chọn hợp nhất hai cụm trong phân cụm phân cấp, $\prod_{k=1}^m P(C_k)$ là không đổi đối với bất kì cặp cụm nào. Do vậy, cho hai cụm C_1, C_2 , khoảng cách giữa chúng được đo bằng $d(C_1, C_2) = -\log \frac{P(C_1 \cup C_2)}{P(C_1) \cdot P(C_2)}$. Phương pháp phân cụm phân cấp

xác suất có thể chấp nhận khung phân cụm mờ rộng, nhưng sử dụng khoảng cách này trong mô hình xác suất.

Theo thay đổi chất lượng khi hợp hai cụm, người ta thấy không luôn luôn dẫn đến chất lượng tốt hơn, tức $\frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)}$ có

thể nhỏ hơn 1. Chẳng hạn các hàm Gauss sử dụng trong hình thí dụ; dù hợp nhất hai cụm C_1, C_2 , được cụm phù hợp với phân phối Gauss, nhưng hợp nhất C_3, C_4 có chất lượng kém đi, do không hàm Gauss nào có thể phù hợp với cụm kết quả.



Hình 7.25. Hợp nhất hai cụm đầu sẽ tăng chất lượng cụm; nhưng hợp nhất hai cụm sau sẽ không tăng

Dựa vào quan sát này, lược đồ phân cụm phân cấp có thể bắt đầu với mỗi cụm gồm một đối tượng, rồi hợp nhất hai cụm C_i và

C_i , nếu khoảng cách giữa chúng là âm. Trông mỗi lần lặp, cố gắng tìm C_i và C_j sao cho cực đại $\log \frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)}$. Việc lặp tiếp tục khi $\log \frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)} > 0$, tức còn có thể cải thiện việc phân cụm. Giả mã như trong hình.

Thuật toán phân cụm phân cấp xác suất

Đầu vào: bộ dữ liệu với n đối tượng $D = \{o_1, \dots, o_n\}$.

Đầu ra: Phân cấp các cụm.

Phương pháp:

Tạo cụm đối với mỗi đối tượng $C_i = \{o_i\}$, $i = 1..n$;

For $i = 1$ to n

Phát hiện cặp cụm C_i, C_j sao cho $(C_i,$

$$C_j) = \arg \max_{i \neq j} \log \frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)};$$

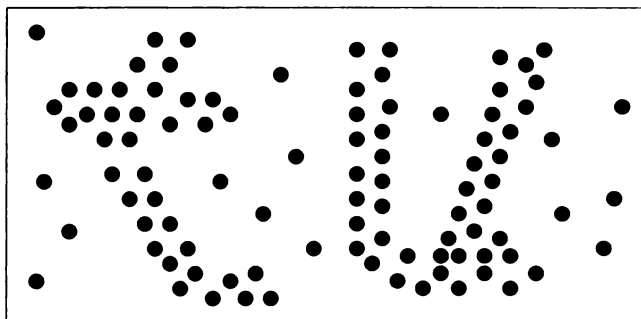
If $\log \frac{P(C_i \cup C_j)}{P(C_i) \cdot P(C_j)} > 0$ then hợp nhất C_i và C_j ;

Else stop.

7.3.4. Các phương pháp dựa trên mật độ

Các phương pháp phân lớp và phương pháp phân cấp được thiết kế để tìm ra các cụm hình cầu. Chúng khó dùng cho các dạng bất kì, chẳng hạn như hình chữ S hay hình ovan. Với dữ liệu như vậy, chúng xác định các miền lỗi không chính xác, có nhiều dữ liệu ồn và dữ liệu lẻ.

Để phát hiện các cụm có hình tùy ý, người ta có thể mô hình hóa các cụm như miền đặc trong không gian dữ liệu, tách nhau nhờ các miền thưa. Đó là chiến lược chính của các *phương pháp phân cụm dựa trên mật độ*, cho phép phát hiện các cụm có hình dạng không như hình cầu.



Hình 7.26. Các cụm có hình bất kì, khó sử dụng phương pháp phân hoạch, phân cấp

7.3.4.1. Phân cụm dựa trên mật độ theo các miền nối với mật độ cao

Mật độ của đối tượng o có thể được đo bởi số các đối tượng sát o . Phương pháp DBSCAN¹ phát hiện các *đối tượng lõi*, tức các đối tượng có các láng giềng dày đặc. Nối đối tượng lõi với các đối tượng láng giềng sẽ tạo nên một cụm.

Liên quan đến việc định lượng tính láng giềng của một đối tượng, người dùng sử dụng tham số $\epsilon > 0$ để xác định bán kính của vùng láng giềng. Lận cận ϵ của đối tượng o là không gian có tâm o , bán kính ϵ .

Tùy theo tham số ϵ , *mật độ của vùng láng giềng* được đo bằng số các đối tượng trong vùng láng giềng. Để xác định vùng láng giềng dày hay thưa, DBSCAN sử dụng tham số m , xác định ngưỡng đặc của miền dày. Một đối tượng là *đối tượng lõi* nếu đối tượng này chứa ít nhất m đối tượng. Các đối tượng lõi là quan trọng đối với miền dày.

Với tập các đối tượng D , người ta có thể xác định các đối tượng lõi ứng với ϵ và m . Nhiệm vụ phân cụm được đưa về việc sử dụng các đối tượng lõi và các láng giềng để tạo nên các miền dày; miền dày là cụm. Đối với đối tượng p và q , có khái niệm *đến được trực tiếp theo mật độ*.

¹ DBSCAN: Density - based Spatial Clustering of Applications with noise: phân cụm không gian dựa trên mật độ của các ứng dụng có ồn.

Định nghĩa: Đối tượng p là đến được trực tiếp theo mật độ từ đối tượng q , ứng với ϵ và m , nếu p thuộc vào lân cận ϵ của q .

Sử dụng quan hệ đến được trực tiếp theo mật độ, đối tượng lõi có thể kéo các đối tượng khác vào cùng một miền dày.

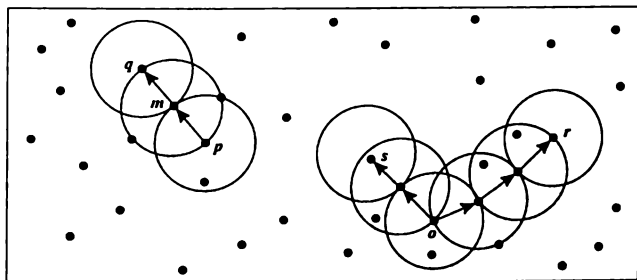
Từ miền dày nhỏ, người ta tạo nên miền dày lớn hơn. Khi p là đến được theo mật độ từ q , với ϵ và m , nếu có dãy $p_1, \dots, p_n$, sao cho $p_1 = q$, $p_n = p$, $p_i \rightarrow p_{i+1}$ là đến được trực tiếp theo mật độ từ p_i . Quan hệ này không đối xứng. Nếu cả o_1 và o_2 là đối tượng lõi và o_1 đến được từ o_2 , thì o_2 là đến được theo mật độ từ o_1 . Tuy nhiên, nếu chỉ o_2 là đối tượng lõi, thì o_1 có thể là đến được theo mật độ từ o_2 , nhưng ngược lại chưa đúng.

Để nối các đối tượng lõi và các láng giềng của chúng vào một miền dày, DBSCAN dùng khái niệm tính nối được theo mật độ.

Định nghĩa: Hai đối tượng p_1, p_2 là nối nhau theo mật độ, với ϵ và m , nếu có đối tượng q sao cho cả p_1 và p_2 là đến được theo mật độ từ q , với ϵ và m .

Không như tính đến được theo mật độ, tính nối được theo mật độ là quan hệ tương đương. Nếu o_1 và o_2 , o_2 và o_3 là nối được theo mật độ, thì o_1, o_3 cũng nối được theo mật độ.

Về tính đến được theo mật độ và nối được theo mật độ: Xét hình vẽ với ϵ là bán kính hình tròn và lấy $m = 3$.



Hình 7.27. Tính đến được theo mật độ và tính nối được theo mật độ trong phân cụm theo mật độ

Các điểm o, p, m, r là các đối tượng lỗi, do chúng ở trong vùng láng giềng có ít nhất 3 đối tượng. Điểm q là đến được trực tiếp theo mật độ từ m . Đối tượng m là đến được trực tiếp theo mật độ từ p và ngược lại. Đối tượng q đến được theo mật độ từ p , nhưng không trực tiếp. Tuy nhiên, p không đến được theo mật độ từ q , do q không là đối tượng lỗi. Tương tự, đối tượng r và s là đến được theo mật độ từ o , và o là đến được theo mật độ từ r . Do vậy, o, r, s là nối được theo mật độ.

Người ta sử dụng bao đóng nối theo mật độ để phát hiện các miền nối dày như các cụm. Mỗi tập đóng là một *cụm dựa trên mật độ*.

Định nghĩa: Tập con $C \subseteq D$ là một cụm nếu (i) đối với hai đối tượng o_1, o_2 bất kì thuộc C , chúng là nối nhau theo mật độ; (ii) không có đối tượng o_1 thuộc C và o_2 thuộc $D - C$, mà o_1 và o_2 là nối nhau theo mật độ.

Liên quan đến cách DBSCAN phát hiện các cụm, ban đầu tất cả các đối tượng trong bộ dữ liệu d là đánh dấu “chưa xét”. Nó chọn ngẫu nhiên đối tượng chưa xét p , đánh dấu p là “đã xét” và kiểm tra láng giềng ϵ có ít nhất m đối tượng không. Nếu không, p được đánh dấu là “điểm ồn”. Ngược lại, đã tạo được cụm C mới cho p , và đưa các đối tượng trong láng giềng ϵ vào tập ứng viên, N . DBSCAN lặp lại việc bổ sung vào C các đối tượng trong N , mà không thuộc vào bất kì cụm nào. Đối tượng p_2 trong N đang mang nhãn “chưa xét” sẽ được đánh dấu là “đã xét” và xét các lân cận ϵ của p_2 . Nếu p_2 có ít nhất m đối tượng láng giềng ϵ , các đối tượng này được bổ sung vào N . DBSCAN tiếp tục bổ sung đối tượng vào C cho đến khi C không nói thêm được, tức N là rỗng. Lúc này C là đầy và là đầu ra.

Để phát hiện cụm tiếp theo, DBSCAN chọn ngẫu nhiên đối tượng chưa xét còn lại. Quá trình phân cụm tiếp tục cho đến khi tất cả các đối tượng được xét. Nếu sử dụng chỉ mục không gian, độ phức tạp tính toán của DBSCAN là $O(n \log n)$, trong đó n là số đối tượng của cơ sở dữ liệu. Nếu không, độ phức tạp là $O(n^2)$. Với tham số ϵ và m thích hợp, thuật toán là hiệu quả để phát hiện các cụm hình dạng bất kì.

Thuật toán DBSCAN, phân cụm dựa trên mật độ.

Đầu vào (i) bộ dữ liệu D có n đối tượng; (ii) tham số bán kính cụm ϵ ; (iii) ngưỡng dày m .

Đầu ra: tập các cụm dựa trên mật độ.

Phương pháp:

Đánh dấu tất cả các đối tượng là "chưa xét";

Do

 Chọn ngẫu nhiên đối tượng chưa xét p ;

 Đánh dấu p là "đã xét";

 If lân cận ϵ của p có ít nhất m đối tượng

 Tạo cụm C mới; bổ sung p vào C ;

 Lấy N là tập các đối tượng lân cận ϵ của p ;

 For mỗi điểm p_2 trong N

 If p_2 là "chưa xét"

 Đánh dấu p_2 là "đã xét";

 If lân cận ϵ của p_2 có ít nhất m đối tượng

 Bổ sung các đối tượng này vào N ;

 If p_2 không là thành viên của cụm nào, bổ sung p_2 vào C ;

 End for;

 Xuất C ;

 Else đánh dấu p là "điểm ồn";

Until không còn đối tượng "chưa xét".

Dù DBSCAN có thể phân cụm các đối tượng với ϵ và m , nó gây khó cho người dùng về việc chọn giá trị tham số để được các cụm chấp nhận được. Đây cũng là vấn đề đối với nhiều thuật toán phân cụm. Việc thiết lập tham số theo kinh nghiệm nên khó tất định, đặc biệt khi có nhiều chiều dữ liệu. Hầu hết các thuật toán nhạy với giá trị tham số: chỉ sai khác nhỏ có thể dẫn đến các cụm khác nhau. Ngoài ra dữ liệu thực tế thường thiên lệch, cấu trúc cụm nội tại không thể đặc trưng bằng tập các tham số mật độ tổng quát.

Các cụm dựa trên mật độ là đơn điệu, theo các ngưỡng. Trong DBSCAN, đối với giá trị m cố định và hai tham số $\varepsilon_1 < \varepsilon_2$, cụm C_1 ứng với ε_1 và m cần là tập con của cụm C_2 , ứng với ε_2 và m . Điều này có nghĩa rằng nếu hai đối tượng thuộc cụm theo mật độ, chúng cần thuộc cụm có tham số mật độ thấp hơn.

7.3.4.2. Thứ tự các điểm để định danh cấu trúc phân cụm

Để khắc phục những khó khăn về một bộ tham số tổng quát trong phân tích phân cụm, người ta đề xuất phương pháp phân tích cụm OPTICS¹, nó không cho kết quả phân cụm theo cách tường minh, mà cho *thứ tự các cụm*.

Định nghĩa: Thứ tự các cụm là danh sách của tất cả các đối tượng cần phân tích.

Thuật toán còn biểu diễn *cấu trúc phân cụm dựa trên mật độ* của dữ liệu. Các đối tượng trong cụm dày hơn được liệt kê sát nhau hơn trong việc xếp thứ tự cụm. Việc sắp xếp này là tương đương với phân cụm dựa trên mật độ với dải rộng các tham số. Vậy OPTICS không đòi hỏi người dùng xác định ngưỡng mật độ. Thứ tự cụm có thể dùng để (i) trích thông tin phân cụm cơ bản, như các tâm cụm, cụm hình dáng bất kì; (ii) rút ra cấu trúc phân cụm theo kinh nghiệm; (iii) hiển thị việc phân cụm.

Để đồng thời kiến thiết phân cụm khác nhau, các đối tượng được xử lí theo trật tự xác định. Trật tự này lựa chọn một đối tượng đến được theo mật độ, với giá trị ε nhỏ nhất, để các cụm với mật độ cao hơn, tức ε thấp hơn, sẽ được kết thúc trước tiên. Theo ý tưởng này, OPTICS cần hai thông tin quan trọng trên mỗi đối tượng:

- *Khoảng cách lõi* của đối tượng p là giá trị ε nhỏ nhất để lân cận ε của p có ít nhất m đối tượng. Tức là ε là ngưỡng khoảng cách tối thiểu để p là đối tượng lõi. Nếu p không là

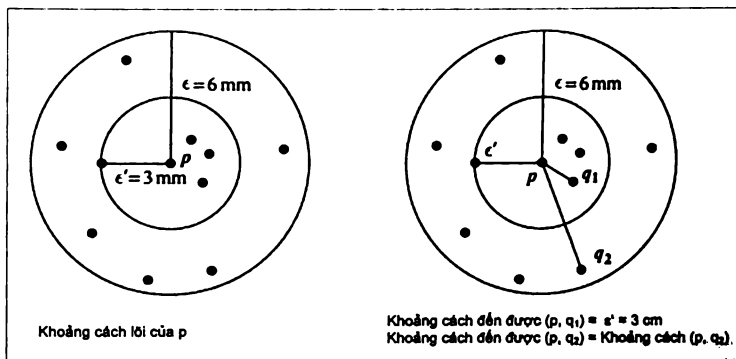
¹ OPTICS: Ordering Points to Identify the Clustering Structure: thứ tự các điểm để định danh cấu trúc phân cụm.

đối tượng lõi, với ϵ và m , người ta không xác định được khoảng cách lõi của p ;

- Khoảng cách đến được đối tượng p từ q là giá trị bán kính tối thiểu để p là đến được theo mật độ từ q . Theo định nghĩa về đến được theo mật độ, q cần là đối tượng lõi, p cần là láng giềng của q . Do vậy, khoảng cách đến được p từ q là $\max \{\text{khoảng cách lõi}(q), \text{khoảng cách}(p, q)\}$. Nếu q không là đối tượng lõi, với ϵ và m , khoảng cách đến được p từ q là không xác định.

Một đối tượng p có thể đến được trực tiếp từ nhiều đối tượng lõi. Do vậy, p có thể có nhiều khoảng cách đến được, ứng với các đối tượng lõi khác nhau. Khoảng cách đến được nhỏ nhất của p có ý nghĩa, do nó cho biết đường ngắn nhất để p nối với cụm dày.

Về khoảng cách lõi và khoảng cách đến được: Xét hình với mô tả khái niệm hai khoảng cách. Giả sử $\epsilon = 6$ mm và $m = 5$. Khoảng cách lõi của p là khoảng cách ϵ' giữa p và bốn đối tượng dữ liệu gần nhất từ p . Khoảng cách đến được của q_1 từ p là khoảng cách lõi của p , tức $\epsilon' = 3$ mm, do nó lớn hơn khoảng cách Oclit từ p đến q_1 . Khoảng cách đến được của q_2 đến p là khoảng cách Oclit từ p đến q_2 , do nó lớn hơn khoảng cách lõi của p .

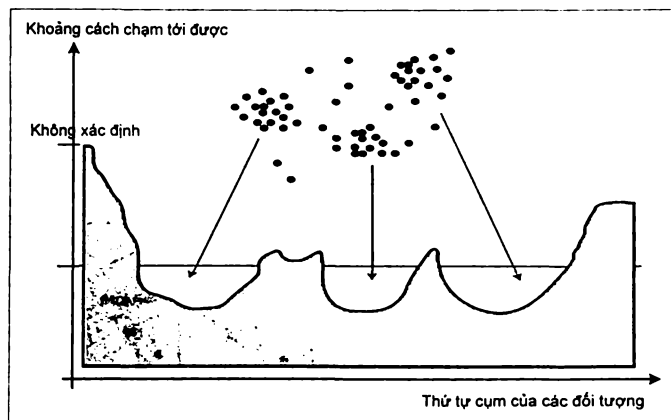


Hình 7.28. Thuật ngữ dùng trong OPTICS

OPTICS tính thứ tự của tất cả các đối tượng trong cơ sở dữ liệu đã biết và đối với mỗi đối tượng trong cơ sở dữ liệu này, r

lưu khoảng cách lỗi và khoảng cách đến được phù hợp. Nó quản lý danh sách, gọi là *hạt giống thứ tự*, để tạo ra thứ tự. Các đối tượng trong hạt giống là sắp xếp theo khoảng cách đến được, từ các đối tượng lỗi gần nhất tương ứng, tức theo khoảng cách đến được nhỏ nhất của mỗi đối tượng.

OPTICS bắt đầu từ đối tượng bất kì trong cơ sở dữ liệu đầu vào làm đối tượng xem xét p . Nó tìm các lân cận ϵ của p , xác định khoảng cách lỗi, và cho khoảng cách đến được là *chưa xác định*. Đối tượng đang xét p được xuất ra. (i) Nếu p không là đối tượng lỗi OPTICS chuyển sang đối tượng tiếp trong danh sách *hạt thứ tự* hay trong cơ sở dữ liệu đầu vào nếu danh sách này rỗng. (ii) Nếu p là đối tượng lỗi, thì đối với mỗi đối tượng q trong lân cận ϵ của p OPTICS cập nhật khoảng cách đến được của nó từ p ; và bổ sung q vào danh sách hạt thứ tự, nếu q chưa được xử lí. Việc này tiếp tục lặp cho đến khi dùng hết đầu vào và hạt thứ tự rỗng.



Hình 7.29. Xếp thứ tự cụm trong OPTICS

Thứ tự cụm của một bộ dữ liệu có thể thể hiện qua đồ họa, để hiển thị và hiểu cấu trúc phân cụm trong bộ dữ liệu. Hình cho thấy đồ thị về *tính đến được* cho bộ dữ liệu 2D, thể hiện tổng quan về cách mà dữ liệu được cấu trúc và phân cụm. Các đối tượng dữ liệu được vẽ theo thứ tự phân cụm, trực hoành, ứng với khoảng cách đến được, trực tung. Ba vết lõm Gauss trên đồ thị thể hiện ba

cụm trong bộ dữ liệu. Cũng có các phương pháp dùng cho các cấu trúc phân cụm trên dữ liệu nhiều chiều.

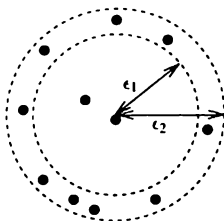
Cấu trúc của thuật toán OPTICS là đơn giản hơn, so với DBSCAN. Hai thuật toán có cùng độ phức tạp thời gian. Độ phức tạp này là $O(n \log n)$ nếu sử dụng chỉ mục không gian, và $O(n^2)$ ngược lại.

7.3.4.3. Phân cụm dựa trên hàm phân phối mật độ

Ước lượng mật độ là điều cốt lõi trong các phương pháp phân cụm dựa trên mật độ. Phương pháp DENCLUE¹ dựa trên tập các hàm phân phối mật độ.

Định nghĩa: *Ước lượng mật độ, trong xác suất và thống kê, là ước lượng về hàm mật độ xác suất không quan sát được, dựa trên tập đối tượng đã quan sát.*

Trong ngữ cảnh phân cụm dựa trên mật độ, hàm mật độ xác suất không quan sát được là phân phối thực của lực lượng của tất cả các đối tượng cần phân tích. Bộ dữ liệu đã quan sát được xem như mẫu ngẫu nhiên từ lực lượng đó.



Hình 7.30. Trong tiếp cận có bán kính ϵ , người ta tăng ϵ để tăng ước lượng mật độ

Trong phương pháp DBSCAN và OPTICS, mật độ được tính theo số các đối tượng láng giềng, xác định theo bán kính ϵ . Mật độ được ước lượng như vậy sẽ rất nhạy với giá trị ϵ . Để khắc phục điều này, người ta dùng *ước lượng mật độ nhân*, là tiếp cận ước

¹ DENCLUE: Density - based Clustering; phân cụm dựa trên mật độ.

lượng mật độ phi tham số, theo thống kê. Ý tưởng của ước lượng mật độ nhân là đơn giản. Người ta xem đối tượng được quan sát như một chỉ số của mật độ xác suất cao trong vùng bao quanh này. Mật độ xác suất tại một điểm phụ thuộc vào các khoảng cách từ điểm này đến các đối tượng được quan sát.

Một cách hình thức, lấy $x_1, \dots, x_n$ là độc lập và mẫu phân phối đồng nhất của biến ngẫu nhiên f . *Xấp xỉ mật độ nhân của hàm mật độ xác suất* là $\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)$; trong đó $K()$ là nhân và h là băng rộng, dùng như tham số làm trơn. *Nhân* là hàm mô hình hóa ảnh hưởng của điểm mẫu trong các điểm láng giềng của nó. Về kỹ thuật, $K()$ là hàm khả tích giá trị thực không âm, thỏa mãn $\int_{-\infty}^{\infty} K(u)du = 1$; $K(-u) = K(u)$ đối với mọi giá trị u . Nhân thường được sử dụng là hàm chuẩn Gauss với trung bình 0 và phương sai 1:

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}.$$

DENCLUE có vài ưu điểm. Nó được xem như tổng quát hóa của vài phương pháp phân cụm, như các tiếp cận liên kết đơn và DBSCAN. Ngoài ra nó không thay đổi theo dữ liệu ồn. Ước lượng mật độ nhân có thể giảm ảnh hưởng của ồn nhờ phân phối thống nhất các ồn vào dữ liệu đầu vào.

7.3.5. Các phương pháp dựa vào lưới

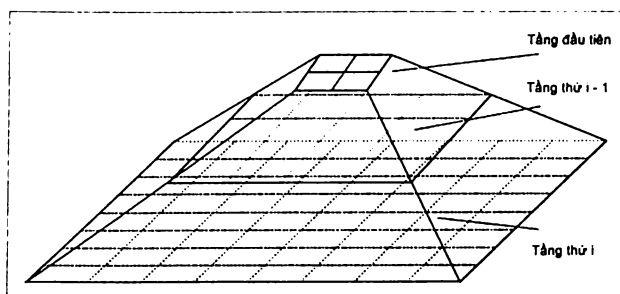
Các phương pháp phân cụm đã xét trên dựa vào dữ liệu; chúng phân hoạch tập các đối tượng và phỏng theo phân phối của các đối tượng trong không gian chứa. Khác với các phương pháp dựa vào dữ liệu là *phương pháp phân cụm dựa vào lưới*, theo tiếp cận không gian, phân hoạch không gian nhúng thành các *tế bào*, độc lập với phân phối của dữ liệu đầu vào.

Định nghĩa: Tiếp cận phân cụm dựa trên lưới là tiếp cận sử dụng cấu trúc dữ liệu lưới với nhiều mức phân giải.

Tiếp cận lưới chia không gian đối tượng thành số hữu hạn các tế bào, tạo nên cấu trúc lưới, mà trên đó người ta thực hiện các phép toán phân cụm. Ưu điểm chính của tiếp cận này là thời gian xử lý nhanh, độc lập với số đối tượng dữ liệu, phụ thuộc chỉ vào số các tế bào trong mỗi chiều trong không gian chia nhỏ.

7.3.5.1. Lưới thông tin thống kê

Kỹ thuật phân cụm theo lưới đa phân giải STING¹ chia miền không gian nhúng của các đối tượng đầu vào thành các tế bào hình chữ nhật. Không gian này được tách theo cách đệ quy, phân cấp. Một vài mức của các tế bào chữ nhật ứng với các mức phân giải khác nhau, tạo nên cấu trúc phân cấp: mỗi tế bào mức cao được phân hoạch, tạo nên số các tế bào ở mức thấp hơn. Thông tin thống kê được xem như các thuộc tính trong mỗi tế bào lưới, như trung bình, giá trị cực đại, cực tiểu, được tính trước và lưu lại như các *tham số thống kê*. Các tham số thống kê này dùng để xử lý câu hỏi và cho các nhiệm vụ phân tích dữ liệu.



Hình 7.31. Cấu trúc phân cấp trong phân cụm STING

Hình vẽ cho thấy cấu trúc phân cấp của phân cụm STING. Các tham số thống kê của các tế bào mức cao được tính theo các tham số mức thấp. Các tham số này gồm (i) tham số độc lập với thuộc tính, *số đếm*; (ii) các tham số phụ thuộc vào thuộc tính, *trung bình, phương sai chuẩn, min, max*; (iii) dạng *phân phối* mà các giá trị

¹ STING: Statistical Information Grid: lưới thông tin thống kê.

thuộc tính trong các tế bào tuân theo, như *chuẩn*, *thống nhất*, hay *không*. Ở đây, thuộc tính là độ đo chọn lọc để phân tích, như *giá*. Khi dữ liệu được đưa vào cơ sở dữ liệu, các tham số *số đếm*, *trung bình*, *phương sai chuẩn*, *min*, *max* của các tế bào mức dưới được tính trực tiếp từ dữ liệu. Giá trị của *phân phối* có thể do người dùng gán, nếu đã biết dạng phân phối, hay lấy từ thử nghiệm giả thuyết như thử nghiệm χ^2 . Dạng phân phối của tế bào mức cao hơn được tính theo đa số của các dạng phân phối của các tế bào mức thấp hơn ứng với nó, với quá trình lọc ngưỡng. Nếu phân phối của các tế bào mức thấp hơn không phù hợp nhau và không thử nghiệm ngưỡng được, dạng phân phối của mức trên là *không*.

Liên quan đến cách dùng thông tin thống kê để trả lời câu hỏi, người ta thấy các tham số thống kê có thể dùng theo cách dựa trên lưới, trên xuống. (i) Tầng trong cấu trúc phân cấp được xác định từ để bắt đầu quá trình trả lời câu hỏi. Tầng này chứa một ít các tế bào. Đối với mỗi tế bào trong tầng hiện tại, người ta tính khoảng tin cậy, hay ước lượng dải xác suất, phản ánh sự thích hợp của tế bào đối với câu hỏi. Các tế bào không thích hợp bị loại bỏ khỏi quá trình xem xét về sau. Việc xử lý của mức thấp tiếp theo chỉ xét các tế bào thích hợp. (ii) Quá trình này lặp lại cho đến khi đến tầng cuối cùng. (iii) Lúc này, nếu thỏa mãn đặc tả câu hỏi, những miền gồm các tế bào thích hợp được trả về. Ngược lại, dữ liệu thuộc các tế bào thích hợp được tìm kiếm và xử lý tiếp, cho đến khi thỏa mãn câu hỏi.

Tính chất hay của STING là đưa đến kết quả phân cụm của DBSCAN nếu mức mịn đạt 0, tức về phía dữ liệu mức rất thấp. Nói khác đi, dùng thông tin về số đếm và kích thước tế bào, các cụm theo mật độ có thể xấp xỉ như tiếp cận của STING. Vậy STING cũng có thể xem là phương pháp phân cụm dựa trên mật độ.

Liên quan đến ưu điểm của STING, người ta thấy (i) tính toán dựa trên lưới là *độc lập với câu hỏi*, do thông tin thống kê lưu trong mỗi tế bào thể hiện thông tin tóm tắt của dữ liệu trong tế bào lưới, *độc lập với câu hỏi*; (ii) cấu trúc lưới thuận tiện với xử lý song song

và cập nhật gia tăng; (iii) hiệu quả của phương pháp này là quan trọng: STING quét cơ sở dữ liệu một lần, để tính các tham số thống kê của các tế bào, nên độ phức tạp thời gian là $O(n)$; trong đó n là số đối tượng. Sau khi tạo được cấu trúc phân cấp, thời gian xử lý câu hỏi là $O(g)$; trong đó g là số các tế bào lưới tại mức thấp nhất, thường nhỏ hơn n .

Do STING sử dụng tiếp cận đa phân giải để phân tích cụm, chất lượng của phân cụm phụ thuộc vào mức độ mịn của mức thấp nhất của cấu trúc lưới. Nếu mức hạt là mịn, giá xử lý sẽ tăng lên; tuy nhiên nếu mức hạt là thô, chất lượng phân tích cụm sẽ giảm. Ngoài ra, STING không để ý đến mối quan hệ không gian giữa các tế bào láng giềng và tế bào con để kiến thiết tế bào cha. Các hình dạng của cụm kết quả có biên cụm là ngang hay dọc, chứ không xiên. Điều này có thể giảm chất lượng và độ chuẩn xác của các cụm, mặc dù thời gian xử lý nhanh.

7.3.5.2. Phương pháp phân cụm không gian con như *Apriori*

Một đối tượng dữ liệu thường có hàng chục thuộc tính; nhiều thuộc tính có thể không thích hợp. Các giá trị thuộc tính có thể thay đổi đáng kể. Các nhân tố này gây khó cho việc định vị cụm trong toàn không gian dữ liệu. Do vậy việc tìm kiếm các cụm trong *không gian con* của dữ liệu là việc làm có ý nghĩa.

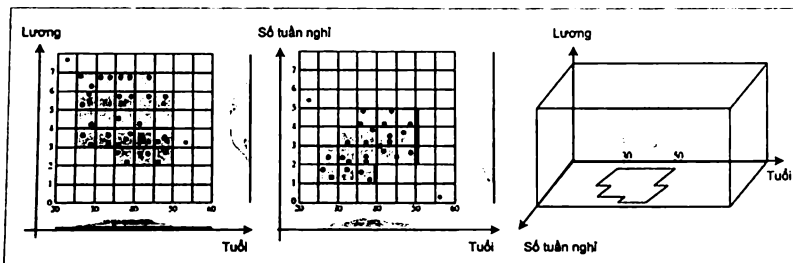
Việc phát hiện nhóm không tầm thường của các bệnh nhân mà các thuộc tính, hay hầu hết thuộc tính, đều thỏa mãn là không thực tế. Đối với bệnh nhân cúm, các thuộc tính *tuổi, giới, công việc* có dải giá trị rộng. Do vậy, chúng gây khó cho việc phát hiện các cụm trong toàn thể không gian. Người ta tìm trong các không gian con, tìm cụm các bệnh nhân tương tự trong không gian ít chiều hơn, tức chẳng hạn tương tự nhau về thuộc tính *sốt cao, ho, tuổi* 3 đến 16.

CLIQUE¹ là phương pháp phân cụm dựa trên lưới đơn giản, để phát hiện các cụm dựa trên mật độ trong không gian con. Nó phân hoạch mỗi chiều thành các khoảng không chồng nhau, rồi phân hoạch toàn bộ không gian nhúng của đối tượng dữ liệu thành các tế bào. Nó sử dụng ngưỡng mật độ để biết các tế bào *dày* hay *thưa*. Một tế bào là *dày* nếu số các đối tượng ứng với nó vượt quá ngưỡng mật độ.

Chiến lược chính của CLIQUE để xác định không gian tìm kiếm ứng viên là sử dụng tính đơn điệu của các tế bào *dày*, ứng với chiều. Điều này theo *tính chất Apriori* dùng trong khai phá luật kết hợp và mẫu phổ biến. Trong ngữ cảnh của các cụm trong không gian con, tính đơn điệu phát biểu như sau.

Định nghĩa: Tính chất Apriori trong CLIQUE là một tế bào k chiều c có thể có ít nhất m điểm chỉ nếu mọi chiều $(k - 1)$ chiều của c , tức tế bào trong không gian con $(k - 1)$ chiều, có ít nhất m điểm.

Trong hình, không gian nhúng dữ liệu có ba chiều *tuổi*, *lương*, *ngày nghỉ*. Xét tế bào 2D tạo bởi *tuổi* và *lương*, có m điểm chỉ nếu chiều của tế bào này trên mọi chiều, tức *tuổi*, *lương*, đều có ít nhất m điểm.



Hình 7.32. Minh họa tính chất Apriori, mức độ dày ứng với mỗi chiều là giao nhau, đảm bảo không gian tìm kiếm ứng viên với mức độ dày ở chiều cao hơn

CLIQUE phân cụm theo hai bước. (i) Đầu tiên nó phân hoạch không gian dữ liệu d chiều thành các phần hình vuông không

¹ CLIQUE: Clustering in quest: phân cụm trong tìm kiếm.

chồng nhau; đặt tên các phần dày trong đó. Để làm điều này, nó phân hoạch mỗi chiều thành các khoảng, đặt tên cho khoảng có ít nhất m điểm; m là ngưỡng dày. Rồi CLIQUE lặp lại việc nối hai tế bào dày k chiều, c_1 và c_2 , trong không gian con $(D_{i1}, \dots, D_{ik})$ và $(D_{j1}, \dots, D_{jk})$, nếu $D_{i1} = D_{j1}, \dots, D_{ik-1} = D_{jk-1}$, và c_1 và c_2 chia sẻ cùng các khoảng trong các chiều này. Phép nối tạo ứng viên $(k+1)$ chiều mới, c , trong không gian con $(D_{i1}, \dots, D_{ik-1}, D_{ik}, D_{jk})$. Nó kiểm tra số các điểm trong c với ngưỡng dày. Việc lặp kết thúc khi không tạo được ứng viên nào, hay không có tế bào ứng viên nào đủ dày. (ii) Trong bước sau, CLIQUE sử dụng các tế bào dày trong mỗi không gian con để hợp thành các cụm, với hình dạng bất kì. Ý tưởng là ứng dụng nguyên tắc *độ dài mô tả tối thiểu* MDL, dùng *vùng cực đại* phủ lên các tế bào dày.

Định nghĩa: Vùng cực đại là siêu hình chữ nhật với mỗi tế bào thuộc vùng này là dày, và vùng này không mở rộng bất kì chiều nào trong không gian con.

Việc phát hiện mô tả tốt nhất cho một cụm, nói chung là bài toán NP khó. Vậy CLIQUE chấp nhận tiếp cận tham lam đơn giản; nó bắt đầu từ tế bào dày bất kì, phát hiện vùng cực đại phủ tế bào này, rồi xét các tế bào khác chưa được phủ. Phương pháp tham lam kết thúc khi đã phủ tất cả các tế bào dày.

Liên quan đến hiệu quả, người ta thấy CLIQUE tự động phát hiện không gian con của chiều cao nhất, sao cho các cụm mật độ cao xuất hiện trong không gian con này. Nó không nhạy với thứ tự của các đối tượng đầu vào và không cần giả thiết về phân phối dữ liệu chính tắc nào. Nó mở rộng tuyến tính theo kích thước của đầu vào, và mở rộng tốt khi số chiều trong dữ liệu tăng lên. Tuy nhiên, phân cụm có ý nghĩa tùy vào việc điều chỉnh thích hợp kích thước của lưới và ngưỡng dày; ở đây đang xét cấu trúc lưới tĩnh. Trong thực tế, việc này có khó khăn, do kích thước lưới và ngưỡng dày dùng cho tổ hợp các chiều trong bộ dữ liệu. Vậy, chuẩn xác của kết quả phân cụm kém đi, bù cho việc đơn giản hóa phương pháp. Ngoài ra, đối với vùng dày đã biết, tất cả chiều của vùng

này trên không gian ít chiều hơn sẽ là dày. Điều này gây nên chồng chéo trong số các vùng được thông báo là dày. Khó khăn khác liên quan đến phát hiện các cụm với mật độ khá khác nhau trong các không gian con có chiều khác nhau.

Người ta đã mở rộng tiếp cận này theo cách đơn giản. Chẳng hạn người ta xem lưới như tập các khối cố định. Thay vì dùng khối cố định đối với mỗi chiều, có thể dùng chiến lược theo dữ liệu, thích nghi, để xác định động các khối đối với mỗi chiều, dựa trên thống kê phân phối dữ liệu. Hoặc thay vì sử dụng ngưỡng, người ta sử dụng độ đo thông tin để đo chất lượng của cụm không gian con.

7.3.6. Đánh giá việc phân cụm

Đánh giá phân cụm theo tính khả thi của phân tích cụm trên bộ dữ liệu và chất lượng của kết quả tạo nên. Các nhiệm vụ chính của đánh giá phân cụm là:

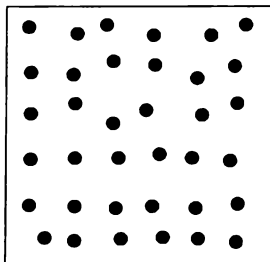
- *Đánh giá khuynh hướng phân cụm.* Đối với bộ dữ liệu đã biết, người ta đánh giá cấu trúc không ngẫu nhiên trong dữ liệu. Việc áp dụng phương pháp phân cụm một cách mù quáng trên bộ dữ liệu vẫn cho ra các cụm, nhưng khiến việc khai phá dữ liệu bị sai. Phân tích cụm trên bộ dữ liệu chỉ có ý nghĩa khi tồn tại cấu trúc không ngẫu nhiên trong dữ liệu;
- *Xác định số cụm trong bộ dữ liệu.* Thuật toán như k trung bình đòi hỏi tham số là số các cụm. Số các cụm được xem như thống kê tóm tắt quan trọng của bộ dữ liệu. Vậy nên ước lượng con số này, trước khi thực hiện thuật toán phân cụm;
- *Đo chất lượng phân cụm.* Sau khi dùng phương pháp phân cụm trên bộ dữ liệu, người ta muốn biết kết quả tốt hay không. Vài phương pháp đo cách các cụm khớp với bộ dữ liệu; hay đo việc các cụm khớp với sự thật. Có phép đo

tính điểm phân cụm, rồi so sánh các kết quả phân cụm trên cùng bộ dữ liệu.

7.3.6.1. Đánh giá khuynh hướng phân cụm

Đánh giá khuynh hướng phân cụm xác định bộ dữ liệu đã biết có cấu trúc không ngẫu nhiên hay không, để các cụm có ý nghĩa. Xét bộ dữ liệu không có cấu trúc không ngẫu nhiên nào, như tập các điểm phân phối thống nhất trong không gian dữ liệu. Cho dù thuật toán phân cụm cho các cụm kết quả, các cụm này là ngẫu nhiên và không mang ý nghĩa nào.

Về phân cụm yêu cầu phân phối không thống nhất của dữ liệu: hình vẽ cho thấy bộ dữ liệu với phân phối thống nhất trong không gian 2D. Dù thuật toán phân cụm vẫn phân hoạch các điểm thành các nhóm, các nhóm này không có ý nghĩa cho ứng dụng, do phân phối thống nhất của dữ liệu.



Hình 7.33. Bộ dữ liệu với phân phối thống nhất trong không gian dữ liệu

Liên quan đến việc đánh giá khuynh hướng của bộ dữ liệu, về trực quan, người ta cố đo xác suất để tạo nên bộ dữ liệu nhờ phân phối dữ liệu thống nhất. Điều này thực hiện được do sử dụng thử nghiệm thống kê đối với tính ngẫu nhiên không gian. Minh họa cho ý tưởng này là thống kê Hopkins.

Định nghĩa: Thống kê Hopkins là thống kê không gian, cho phép thử nghiệm tính ngẫu nhiên không gian của một biến phân phối trong không gian.

Cho bộ dữ liệu D , được xem như mẫu của một biến ngẫu nhiên o . Người ta muốn xác định o ở xa ra sao so với cái theo phân phối thống nhất trong không gian dữ liệu. Người ta tính thống kê Hopkins theo:

1. Lấy mẫu thống nhất n điểm $p_1, \dots, p_n$, từ D . Tức mỗi điểm trong D có cùng xác suất tham gia mẫu này. Đối với mỗi điểm p_i , người ta tìm láng giềng gần nhất của p_i , và x_i là khoảng cách giữa p_i và láng giềng này. Tức $x_i = \min_{v \in D} \{d(p_i, v)\}$;
2. Lấy mẫu thống nhất n điểm từ D , $q_1, \dots, q_n$. Đối với mỗi q_i , người ta tìm láng giềng gần nhất của q_i trong $D - \{q_i\}$, và khoảng cách giữa chúng là y_i . Tức $y_i = \min_{v \in D - \{q_i\}} \{d(q_i, v)\}$;

$$3. \text{ Tính thống kê Hopkins } H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}.$$

Liên quan đến thống kê Hopkins cho biết khả năng bộ dữ liệu D theo phân phối thống nhất trong không gian dữ liệu, người ta thấy nếu D theo phân phối thống nhất, thì $\sum_{i=1}^n x_i$ và $\sum_{i=1}^n y_i$ sẽ gần nhau, nên H sẽ gần 0.5. Tuy nhiên, nếu D có dạng lệch, thì $\sum_{i=1}^n y_i$ sẽ nhỏ hơn đáng kể so với $\sum_{i=1}^n x_i$, nên H gần với 0.

Giả thiết rằng là *giả thuyết đồng nhất*, tức D là phân phối thống nhất và không có cụm mang ý nghĩa. Giả thuyết khác là *giả thuyết không đồng nhất*, tức D không theo phân phối thống nhất, có cụm. Người ta có thể thử nghiệm thống kê Hopkins nhiều lần, sử dụng 0.5 như ngưỡng; tức nếu $H > 0.5$ thì có vẻ D có các cụm có ý nghĩa.

7.3.6.2. Xác định số cụm

Xác định số cụm đúng trong bộ dữ liệu là quan trọng, không chỉ cho thuật toán phân cụm như k trung bình làm tham số, mà số phù hợp sẽ điều khiển mức độ mịn của phân tích cụm. Việc này được xem như việc cân đối giữa *tính nén* và *tính chuẩn xác* trong

phân tích cụm. Xét các trường hợp cực đoan. (i) Sẽ ra sao nếu xem toàn bộ dữ liệu như một cụm. Nó sẽ tăng tính nén của dữ liệu, nhưng không có giá trị phân tích cụm. (ii) Mặt khác, nếu mỗi đối tượng là một cụm, người ta được độ phân giải phân cụm tinh, tức chuẩn xác nhất theo khoảng cách 0 giữa một đối tượng và tâm cụm. Trong vài phương pháp như k trung bình, việc này cần đến chi phí cao. Tuy nhiên, một đối tượng trong một cụm không thể nhận xét về dữ liệu được.

Việc xác định số cụm không dễ dàng, do khái niệm số cụm đúng còn nhập nhằng. Phát hiện đúng số cụm phụ thuộc vào dạng của phân phối và quy mô của bộ dữ liệu, cũng như độ phân giải phân cụm mà người dùng yêu cầu. Có nhiều cách để ước lượng số cụm.

Phương pháp đơn giản nhất là cho số cụm là khoảng $\sqrt{\frac{n}{2}}$, đối với bộ n dữ liệu. Người ta hi vọng mỗi cụm có $\sqrt{2n}$ điểm.

Phương pháp khuỷu tay dựa trên quan sát rằng số cụm có thể giúp giảm tổng của các phương sai trong cụm của mỗi cụm. Lí do là có càng nhiều cụm sẽ cho phép người ta có cụm càng mịn. Tuy nhiên, việc giảm tổng phương sai trong cụm không cần nữa, khi có tạo nên quá nhiều cụm, bởi vì việc tách các cụm đã dày đặc sẽ chỉ giảm chút ít. Như vậy là việc chọn số cụm theo kinh nghiệm được dùng để điều khiển đường cong của tổng các phương sai giữa cụm, ứng với số các cụm.

Về kĩ thuật, cho số $k > 0$, người ta có thể tạo nên k cụm trong bộ dữ liệu nhờ thuật toán phân cụm như k trung bình, và tính tổng các phương sai giữa cụm $v(k)$. Rồi người ta vẽ đường cong $v(k)$ theo k . Điểm quặt đầu tiên, hay điểm quặt có ý nghĩa, của đường cong này ứng với số cụm đúng.

Các phương pháp tiên tiến hơn có thể xác định số cụm theo tiêu chuẩn thông tin hay các tiếp cận lí thuyết thông tin.

Số cụm đúng trong bộ dữ liệu cũng có thể xác định theo *kiểm chứng chéo*, thường sử dụng trong phân lớp: (i) trước tiên chia bộ

dữ liệu D thành m phần; (ii) dùng $m - 1$ phần để xây dựng mô hình phân cụm, và dùng phần còn lại để kiểm tra chất lượng của việc phân cụm; chẳng hạn đối với mỗi điểm trong bộ dữ liệu thử nghiệm, có thể phát hiện tâm gần nhất; (iii) dùng tổng khoảng cách bình phương giữa tất cả các điểm trong bộ thử nghiệm và các tâm gần nhất để đo việc khớp của mô hình phân cụm với dữ liệu thử nghiệm. Đối với $k > 0$ bất kì, người ta lặp lại quá trình m lần, để phân k cụm nhờ sử dụng mỗi lần một phần dữ liệu thử nghiệm. Trung bình của số đo chất lượng được lấy làm số đo chất lượng chung. Người ta có thể so sánh số đo chất lượng chung với các giá trị khác nhau của k , để phát hiện số cụm đúng với bộ dữ liệu.

7.3.6.3. Đo chất lượng phân cụm

Giả sử đã đánh giá được khuynh hướng phân cụm của bộ dữ liệu đã biết. Cũng có thể đoán được số c cụm. Rồi sử dụng thuật toán phân cụm để có các cụm. Vấn đề lúc ấy là các cụm này có tốt không và so sánh việc phân cụm của thuật toán này với thuật toán khác.

Có các phương pháp để đo chất lượng phân cụm. Nói chung các phương pháp này được chia thành hai nhóm, theo khả năng khớp với sự thật. Sự thật là phân cụm lí tưởng, thường do chuyên gia thực hiện. (i) Nếu có sẵn sự thật, người ta dùng các *phương pháp ngoài*, đo khi so sánh kết quả phân cụm với sự thật. (ii) Ngược lại, dùng các *phương pháp nội tại*, ước lượng mức độ tốt của việc phân cụm, thông qua việc xem các cụm tách biệt nhau ra sao. Sự thật có thể được coi là dạng giám sát về các *nhân cụm*. Vậy nên phương pháp ngoài được biết như *phương pháp giám sát*; phương pháp nội tại được biết như *phương pháp không giám sát*.

Các phương pháp ngoài: khi có sẵn sự thật, người ta có thể so sánh nó với phân cụm, để đánh giá việc phân cụm. Vậy cốt lõi của các phương pháp ngoài là gán điểm $Q(C, C_s)$ cho việc phân cụm C , với sự thật C_s . Phương pháp ngoài là hiệu quả ra sao, phụ thuộc nhiều vào việc sử dụng Q .

Nói chung, độ đo Q về chất lượng phân cụm là hiệu quả, nếu thỏa mãn bốn tiêu chí sau:

1. *Đồng nhất của cụm.* Cụm càng linh, việc phân cụm càng tốt. Giả sử sự thật là các đối tượng trong bộ dữ liệu D thuộc về các phạm trù $L_1, \dots, L_n$. Xét việc phân cụm C_1 ; cụm $C \in C_1$ chứa các đối tượng từ hai phạm trù L_i, L_j . Có phân cụm C_2 , như C_1 , ngoại trừ việc C_2 tách thành hai cụm chứa các đối tượng trong L_i, L_j . Phép đo chất lượng cụm Q cần cho C_2 điểm cao hơn, tức $Q(C_2, C_g) > Q(C_1, C_g)$;
2. *Tính đầy đủ của cụm.* Tính chất này đòi hỏi rằng đối với mỗi cụm, nếu bất kì hai đối tượng sự thật thuộc cùng phạm trù, thì chúng được gán cùng cụm. Xét việc phân cụm C_1 , chứa các cụm X, Y , có phần tử thuộc về cùng phạm trù như sự thật. Xét C_2 , như C_1 , ngoại trừ X, Y được xếp cùng cụm trong C_2 . Vậy thì độ đo Q cần cho C_2 điểm cao hơn, tức $Q(C_2, C_g) > Q(C_1, C_g)$;
3. *Túi rách.* Trong nhiều kịch bản thực tế, thường có phạm trù *túi rách*, chứa các đối tượng không thể hợp với các đối tượng khác. Phạm trù ấy thường được gọi là *lộn xộn, khác...* Tiêu chí *túi rách* khẳng định rằng đặt một đối tượng di biến vào một cụm tinh sẽ có hại hơn việc đặt nó vào túi rách. Xét phân cụm C_1 , có cụm $C \in C_1$, mà tất cả các đối tượng trong C , trừ đối tượng o , sự thật thuộc về cùng phạm trù. Có phân cụm C_2 , như C_1 , ngoại trừ o được gán cho cụm $C' \neq C$ trong C_2 , sao cho C' sự thật chứa các đối tượng từ nhiều phạm trù, tức dữ liệu ồn. Nói khác đi, C' trong C_2 là túi rách. Vậy độ đo Q cần cho $Q(C_2, C_g) > Q(C_1, C_g)$;
4. *Bảo toàn cụm nhỏ.* Nếu phạm trù nhỏ được tách thành phần nhỏ trong phân cụm, các phần nhỏ này có vẻ trở thành dữ liệu ồn, và phạm trù nhỏ đó mất khỏi việc phân cụm. Tiêu chí về bảo toàn cụm nhỏ khẳng định rằng việc tách phạm trù nhỏ thành các phần là có hại hơn việc tách phạm trù

lớn thành các phần. Xét trường hợp cực đoan. Cho bộ dữ liệu D có $n + 2$ đối tượng, theo sự thật thì n đối tượng $o_1, \dots, o_n$ thuộc về một phạm trù, và đối tượng o_{n+1} và o_{n+2} thuộc về phạm trù khác. Giả sử phân cụm C_1 có ba cụm $X = \{o_1, \dots, o_n\}$, $Y = \{o_{n+1}\}$ và $Z = \{o_{n+2}\}$. Có phân cụm C_2 , cũng cho ba cụm $X_1 = \{o_1, \dots, o_{n-1}\}$, $Y_2 = \{o_n\}$ và $Z_3 = \{o_{n+1}, o_{n+2}\}$. Nói khác đi, C_1 tách phạm trù nhỏ và C_2 tách phạm trù lớn. Độ đo Q xác định $Q(C_2, C_g) > Q(C_1, C_g)$.

Nhiều độ đo chất lượng phân cụm thỏa mãn vài tiêu chí trong số bốn tiêu chí trên. Người ta dùng độ đo BCubed, đo *độ chính xác* và *độ nhạy* đối với mọi đối tượng trong phân cụm trên bộ dữ liệu đã biết, theo sự thật.

Định nghĩa: Độ chính xác của một đối tượng cho biết bao nhiêu đối tượng khác trong cùng cụm, thuộc về cùng phạm trù như đối tượng này.

Định nghĩa: Độ nhạy của một đối tượng phản ánh bao nhiêu đối tượng cùng phạm trù được gán cùng cụm.

Cho bộ dữ liệu $D = \{o_1, \dots, o_n\}$, có phân cụm C_1 . $L(o_i)$ là phạm trù của o_i , theo sự thật, và $C(o_i)$ là tên cụm của o_i trong C_1 . Đối với hai đối tượng $o_i, o_j, i \neq j$, tính đúng đắn của quan hệ giữa o_i và o_j trong C_1 được xác định đúng $(o_i, o_j) = 1$ nếu $L(o_i) = L(o_j) \Leftrightarrow C(o_i) = C(o_j)$; 0 nếu ngược lại. Xác định độ chính xác p , độ nhạy r theo DCubed là:

$$p = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{o_j: i \neq j, C(o_i)=C(o_j)} \text{đúng}(o_i, o_j)}{\|\{o_j \mid i \neq j, C(o_i) = C(o_j)\}\|}$$

$$r = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{o_j: i \neq j, L(o_i)=L(o_j)} \text{đúng}(o_i, o_j)}{\|\{o_j \mid i \neq j, L(o_i) = L(o_j)\}\|}$$

Các phương pháp nội tại: Khi chưa biết sự thật, người ta cần dùng phương pháp nội tại để đánh giá chất lượng phân cụm. Nói chung các phương pháp nội tại ước lượng việc phân cụm nhờ kiểm tra các cụm được tách ra sao và cô đặc ra sao. Nhiều phương

pháp nội tại có ưu điểm về độ đo tương tự giữa các đối tượng trong bộ dữ liệu.

Hệ số bóng là độ đo như vậy. Cho bộ dữ liệu D có n đối tượng; giả sử D được phân hoạch thành k cụm $C_1, \dots, C_k$. Đối với mỗi đối tượng $o \in D$, người ta tính $a(o)$ như khoảng cách trung bình giữa o và tất cả các đối tượng khác trong cụm mà o đang thuộc. Tương tự $b(o)$ là khoảng cách trung bình tối thiểu từ o đến tất cả các cụm mà o không thuộc. Giả sử $o \in C_i$, người ta tính $a(o)$, $b(o)$ rồi tính được hệ số bóng của o :

$$a(o) = \frac{\sum_{o' \in C_i, o' \neq o} d(o, o')}{|C_i| - 1}$$

$$b(o) = \min_{C_j, 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} d(o, o')}{|C_j|} \right\}$$

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

Giá trị của hệ số bóng thuộc $(-1, 1)$. Giá trị $a(o)$ phản ánh độ dày đặc của cụm có o . Giá trị này càng nhỏ, cụm càng dày đặc. Giá trị $b(o)$ cho biết mức độ o tách biệt với các cụm khác. Giá trị này càng lớn, o càng tách biệt với các cụm khác. Do vậy, khi $s(o)$ tiến đến 1, cụm chứa o là dày và o xa các cụm khác. Khi $s(o)$ âm, tức $b(o) < a(o)$, o gần với các đối tượng cụm khác, hơn là gần đối tượng cùng cụm; nói chung đây là trường hợp nên tránh.

Để đo mức khớp của cụm trong phân cụm, người ta tính hệ số bóng trung bình của tất cả các đối tượng trong cụm. Để đo chất lượng của phân cụm, có thể dùng giá trị hệ số bóng trung bình của tất cả các đối tượng trong bộ dữ liệu. Hệ số bóng và độ đo nội tại khác cũng có thể được dùng trong phương pháp khuỷu tay, để có được số cụm trong bộ dữ liệu, theo cách may rủi, thay cho việc tính tổng các phương sai trong cụm.

7.3.7. Kết luận về phân cụm

- *Cụm* là tập hợp đối tượng dữ liệu *tương tự* nhau trong cùng cụm, *phân biệt* nhau với đối tượng khác cụm. Quá trình nhóm tập các đối tượng trừu tượng, vật lí thành lớp các đối tượng *tương tự* gọi là *phân cụm*;
- Phân tích cụm *ứng dụng* trong thông minh doanh nghiệp, nhận dạng ảnh, tìm kiếm Web, sinh học và an toàn. Phân tích cụm được dùng như công cụ khai phá dữ liệu độc lập, với phân phối dữ liệu, hoặc như bước tiền xử lí cho thuật toán khai phá dữ liệu trên các cụm phát hiện được;
- Phân cụm là lĩnh vực nghiên cứu động trong khai phá dữ liệu. Nó liên quan đến *học phi giám sát* trong học máy;
- Việc phân cụm còn nhiều thách thức. *Yêu cầu điển hình* là (i) việc mở rộng; (ii) khả năng xử lí hình dạng bất kì; (iii) yêu cầu tri thức lĩnh vực tối thiểu khi xác định tham số đầu vào; (iv) khả năng xử lí dữ liệu ồn; (v) phân cụm gia tăng và không nhạy với thứ tự đầu vào; (vi) khả năng phân cụm dữ liệu nhiều chiều; (vii) phân cụm với ràng buộc; (viii) tính sử dụng và diễn giải được;
- Có nhiều thuật toán phân cụm. Chúng được phân lớp theo khía cạnh phân biệt, như (i) tiêu chí phân hoạch; (ii) việc tách cụm; (iii) độ đo tương tự; (iv) không gian phân cụm. Phần trên đã đề cập (i) *các phương pháp phân hoạch*; (ii) *các phương pháp phân cấp*; (iii) *các phương pháp dựa trên mật độ*; (iv) *các phương pháp lưới*;
- *Phương pháp phân hoạch* trước tiên tạo k đoạn, trong đó k là số đoạn kiến thiết. Rồi nó *tái phân phối lặp*, cải thiện việc phân hoạch, nhờ chuyển các đối tượng từ nhóm này sang nhóm khác. Phương pháp phân hoạch điển hình là k trung bình, k tâm đại diện và CLARANS;
- *Phương pháp phân cấp* cho phân rã phân cấp của bộ dữ liệu các đối tượng. Phương pháp này có thể theo (i) nói rộng,

dưới lên; (ii) chia nhỏ, trên xuống, dựa trên cách tạo nên phân rã. Để bù cho việc *tách, nhập* cứng nhắc, chất lượng của nới rộng phân cấp được cải thiện nhờ phân tích liên kết đối tượng tại mỗi phân hoạch phân cấp, như trong Chameleon; hay thực hiện *phân cụm nhỏ*, tức nhóm các đối tượng vào cụm nhỏ, rồi làm việc trên cụm nhỏ với kĩ thuật phân cụm, như trong BIRCH;

- *Phương pháp dựa trên mật độ* dùng khái niệm độ dày. Nó phát triển cụm hoặc theo mật độ của các đối tượng láng giềng, như DBSCAN, hay theo hàm mật độ, như DENCLUE. Phương pháp OPTICS dựa trên mật độ, tạo thứ tự tăng của cấu trúc phân cụm dữ liệu;
- *Phương pháp dựa trên lưới* chia không gian đối tượng thành số hữu hạn các tế bào, tạo nên cấu trúc lưới, rồi thực hiện phân cụm trên cấu trúc lưới. STING là thí dụ điển hình của phương pháp này, dựa trên thông tin thống kê lưu trong các tế bào lưới. CLIQUE là phương pháp phân cụm trong không gian con;
- *Đánh giá phân cụm* có nhiệm vụ đánh giá tính khả thi của phân tích cụm trên bộ dữ liệu và chất lượng của kết quả thu được. Có các nhiệm vụ đánh giá (i) khuynh hướng phân cụm; (ii) xác định số cụm; (iii) đo chất lượng phân cụm.

7.4. Kết luận

Chương trên đã đề cập thuật toán (i) phân lớp; (ii) phân cụm. Hai lớp thuật toán khai phá dữ liệu là lớp tiếp theo lớp thuật toán khai phá luật kết hợp.

Ba tiếp cận phân lớp là (i) phân lớp với cây quyết định; (ii) phân lớp với yếu tố Bayes; (iii) phân lớp với các luật. Bốn tiếp cận phân cụm là (i) phân cụm theo phân hoạch; (ii) phân cụm theo phân cấp; (iii) phân cụm theo mật độ; (iv) phân cụm theo lưới.

Tài liệu tham khảo

1. D. Pyle, *Data Preparation for Data Mining*, Ed. Morgan Kaufmann, 1999
2. Daniel T. Larose, *Data mining methods and models*, Ed. Wiley, 2006
3. David Hand, Heikki Mannila, Padhraic Smyth, *Principles of Data Mining*, Ed. MIT Press, 2001
4. Jiawed Han, Micheline Kamper, *Datamining, concepts and technique*, 3rd ed., Morgan Kaufmann Press, 2012
5. M. J. de Smith, *Statistical Analysis Handbook*, The Winchelsea Press, Winchelsea, UK, 2011
6. Mohammed J. Zaki, Wagner Meira JR., *Datamining and Analysis. Fundamental Concepts and Algorithms*, Universidade Federal de Minas Gerais, Brazil, Ed. Cambridge University Press, 2014.
7. Nong Ye, *The Hanbook of Data Mining*, Ed. Lawrence Erlbaum Associates, 2003
8. Peter Cabena, et. al., *Discovering Data Mining, from concept to implementation*, Ed. Prentice Hall, 1997
9. R. Nisbet, J. Elder, G. Miner, *Hanbook of Statistical Analysis and Data Mining Applications*, Ed. Elservier, 2009.
10. Soumen Chakrabarti, *Mining the Web, discovering knowledge from hypertext data*, Ed. Morgan Kaufmann, 2003
11. Usama M. Fayyad, et. al., *Advances in Knowledge Discovery and Data Mining*, Ed. AAAI Press / The MIT Press, 1996

Index

A

Apriori..... 500, 510, 521, 522, 523,
524, 525, 526, 527, 530, 535, 536,
537, 549, 649, 650
ARIMA .11, 103, 366, 374, 378, 379,
380, 381, 382, 383, 384

B

Biểu đồ điều khiển..... 12, 425, 426,
427, 428, 430, 431, 432, 433, 434,
435, 436, 437, 438, 439, 440, 441,
442, 444, 445, 446, 447, 448, 449,
450, 452, 453, 454, 455, 456, 457,
458, 459

C

Cây quyết định..... 13, 51, 213, 500,
555, 556, 569, 579, 590, 604
Chuẩn bị dữ liệu 6, 8, 9, 20, 36, 47,
49, 55, 83, 85, 109, 187, 188, 194,
195, 196, 198, 199, 202, 203, 205,
206, 208, 209, 210, 211, 212, 214,
215, 216, 220, 222, 271, 418, 421,
485
Chuẩn hóa..... 10, 218, 290
Chuỗi thời gian 11, 52, 85, 100,
101, 102, 103, 118, 122, 362, 363,
364, 365, 366, 367, 369, 370, 371,
372, 373, 374, 375, 378, 379, 380,
383, 384, 385, 386, 387

D

Dữ liệu phạm trù . 83, 91, 212, 218,
220, 273, 305, 315, 321, 387, 396,
629

Dữ liệu số 83, 90, 91, 109, 122,
123, 212, 220, 273, 280, 322, 611,
612, 629, 630
Độ đo..... 6, 7, 65, 73, 110, 113, 115,
116, 124, 133, 144, 145, 200, 201,
202, 204, 220, 276, 277, 278, 320,
541, 544, 546, 547, 560, 561, 568,
585, 588, 613, 628, 658

H

Hiệu dữ liệu 6, 84, 188
Hướng người dùng 5, 39, 40, 41,
42, 134, 145

K

Khoảng cách..... 10, 114, 116, 173,
174, 175, 291, 292, 294, 297, 320,
321, 333, 627, 628, 629, 642, 643
Không chắc chắn .. 12, 48, 196, 210,
459, 460, 461, 470, 472, 474, 481,
485
Khuôn mẫu ... 11, 12, 21, 31, 34, 35,
36, 37, 38, 39, 42, 43, 44, 46, 48,
50, 51, 62, 193, 195, 389, 390,
392, 397, 400, 401, 402, 403, 427,
430, 432, 436, 438, 439, 459, 465,
466, 485, 488, 490, 491, 493, 494,
495, 496, 497, 498

L

Làm sạch dữ liệu 83, 85, 89, 365
Lọc dữ liệu 66, 85, 89, 99
Luật kết hợp.. 12, 13, 16, 44, 51, 56,
63, 67, 68, 74, 75, 402, 487, 493,
510, 511, 512, 515, 517, 518, 519,
520, 521, 526, 527, 540, 541, 544,
548, 549, 550, 650, 661

M

Mạng nơron .. 30, 31, 32, 33, 36, 50,
51, 57, 91, 92, 98, 104, 227, 257,
263, 265, 406, 408, 573, 575, 576,
608

Mô hình hóa.... 8, 9, 11, 12, 20, 35,
36, 39, 43, 52, 83, 86, 88, 89, 90,
91, 97, 99, 100, 101, 103, 104,
105, 106, 107, 109, 121, 133, 157,
161, 188, 194, 195, 196, 197, 198,
206, 207, 209, 210, 211, 212, 214,
215, 216, 217, 219, 221, 222, 223,
224, 249, 258, 259, 260, 261, 264,
286, 298, 300, 304, 305, 306, 307,
315, 317, 320, 321, 351, 363, 366,
371, 374, 378, 385, 390, 393, 394,
397, 399, 403, 408, 409, 413, 414,
418, 419, 421, 422, 459, 461, 482,
492, 493, 497, 515, 607, 632, 633,
637, 646

N

Ngôn ngữ hỏi . 6, 59, 66, 77, 78, 79,
81, 82, 86, 216, 221, 250

P

Phân lớp ... 13, 16, 38, 39, 43, 56, 57,
61, 62, 64, 69, 72, 73, 74, 75, 77,
78, 80, 82, 136, 170, 198, 222,
223, 228, 256, 292, 329, 391, 393,
394, 397, 398, 436, 488, 490, 492,
493, 495, 498, 499, 501, 502, 503,
504, 505, 506, 507, 508, 509, 527,

550, 551, 552, 553, 554, 555,
556, 557, 561, 562, 563, 565, 569,
573, 574, 575, 576, 577, 578, 579,
580, 581, 582, 583, 584, 585, 586,
587, 588, 589, 590, 591, 593, 594,
595, 596, 597, 598, 599, 600, 601,
602, 603, 604, 605, 606, 607, 608,
609, 610, 611, 637, 655, 660, 661

Phân phối dữ liệu . 8, 28, 33, 36, 70,
89, 96, 98, 100, 128, 147, 468,
605, 610, 651, 652, 653, 660

Phân tích cụm 14, 43, 492, 609,
610, 652, 660

Phân tích thành phần chính 106,
129, 130, 132, 133, 398, 399, 492

Phương pháp chọn mẫu .. 230, 231,
234, 235

R

Ra quyết định..... 12, 417

S

Suy luận thống kê 12, 23, 117, 239,
461, 465, 478, 481, 482

T

Tích hợp dữ liệu 47, 84, 221
Tiếp cận Bayes . 23, 29, 30, 462, 471,
472, 475

Thu thập dữ liệu 7, 84, 111, 119,
121, 122, 230, 231, 235, 436, 438,
440, 472, 481, 496, 595

Trung bình động 11, 366, 367,
368, 372

NHÀ XUẤT BẢN

ĐẠI HỌC QUỐC GIA HÀ NỘI

16 Hàng Chuối - Hai Bà Trưng Hà Nội

Giám đốc - Tổng Biên tập: (04)39715011

Quản lý xuất bản: (04)39728806; Fax: (04)39724736

Biên tập: (04)39714896

Kỹ thuật xuất bản: (04)39715013

Chịu trách nhiệm xuất bản: Giám đốc - Tổng Biên tập: TS. PHẠM THỊ TRÂM

Hội đồng nghiệm thu giáo trình
Trường Đại học Khoa học Tự nhiên - ĐHQGHN

Người nhận xét: PGS. TS. NGUYỄN LINH GIANG
PGS. TS. NGUYỄN HÀ NAM

Biên tập: TRỊNH THỊ THU HÀ
Chế bản: NGUYỄN SỸ DƯƠNG
Trình bày bìa: NGUYỄN NGỌC ANH

PHÂN TÍCH THỐNG KÊ VÀ KHAI PHẪ DỮ LIỆU

Mã số: 1K – 03 ĐH2018

In 300 cuốn, khổ 16x24 cm tại Công ty CP in Thương mại Truyền Thông Việt Nam
số 7, ngách 28, ngõ 29 phố Vĩnh Tuy, phường Vĩnh Tuy, Quận Hai Bà Trưng, Hà Nội

Số xuất bản: 661 - 2018/CXBIPH/03 - 79/ĐHQGHN, ngày 01/3/2018

Quyết định xuất bản số: 17 KH - XH/QĐ - NXBĐHQGHN, ngày 20/4/2018

In xong và nộp lưu chiểu năm 2018